# TIME-SCALE MODIFICATION OF SPEECH BASED ON SHORT-TIME FOURIER ANALYSIS

BY

MICHAEL RODNEY PORTNOFF

S.B., S.M., E.E., Massachusetts Institute of Technology
(1973)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

April 1978

Signature of Author.....
Department of Electrical Engineering and Computer Science,
April 20, 1978

Certified by.....
Thesis Supervisor

Accepted by.....
Chairman, Departmental Committee on Graduate Students

# TIME-SCALE MODIFICATION OF SPEECH BASED ON SHORT-TIME FOURIER ANALYSIS

by

Michael Rodney Portnoff

Submitted to the Department of Electrical Engineering and Computer Science, on April 20, 1978, in partial fulfillment of the requirements for the degree of Doctor of Science.

## ABSTRACT

The theoretical basis for representation of a speech signal by its short-time Fourier transform and the application of this representation to the problem of time compression and expansion of speech are developed. A time-frequency representation for linear time-varying systems is discussed and applied to a model for speech production to formulate a quasi-stationary representation for the speech waveform. This representation has the property that simple time scaling of the parameters of the representation corresponds to changing the rate of the speech. Given a real speech signal, short-time Fourier analysis provides a technique for estimating and modifying these parameters. The results of the theoretical analysis are used to design a high-quality speech rate-change system which was simulated on a general-purpose digital minicomputer.

In addition to the application of short-time Fourier analysis to speech, a number of more general results are derived. The problems of representing a signal by its short-time Fourier transform and resynthesizing the signal from its transform are considered. A new synthesis equation is introduced which is sufficiently general to describe apparently different synthesis techniques reported in the literature. It is shown that a class of linear-filtering problems can be represented as the product of the time-varying frequency response of a linear filter and the short-time Fourier transform of the input signal. These results are extended to the discrete (sampled) short-time Fourier transform representation, and computationally efficient algorithms, based on the fast Fourier transform, are developed for implementing discrete short-time Fourier analysis and synthesis.

THESIS SUPERVISOR: Alan V. Oppenheim
TITLE: Professor of Electrical Engineering

ACKNOWLEDGMENT

I would like to express my gratitude to Prof. Alan V. Oppenheim for his guidance and encouragement. He provided the initial motivation for this thesis, and his insight, critical comments, and uncompromising standards have had a major effect in shaping the outcome of this work.

I was fortunate to have had the benefit of advise from Prof. James H. McClellan and Prof. Kenneth N. Stevens who assisted in the supervision of this research. Also, I would like to thank Dr. Dan E. Dudgeon who read and commented on an early version of the chapters on short-time Fourier analysis.

I would like to acknowledge Mr. Gary E. Kopec for several meaningful technical discussions during the early phases of this research. I am also grateful for his willingness to provide friendly and expert advice about the computer system and for contributing many computer programs.

Finally, I would like to express my appreciation to the members, past and present, of the Digital Signal Processing Group and the Speech Communication Research Group who have contributed greatly to making this a meaningful and rewarding experience.

TABLE OF CONTENTS

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 INTRODUCTION

The use of recorded speech as a medium for disseminating information is attractive for many reasons. Recorded speech can be understood by anyone who understands the spoken language. Audio recordings of conversations or meetings not only convey more information (such as intonation, mood, etc.) than written transcripts, but also eliminate the possibility of errors introduced by the transcription process. Moreover, such recordings are more easily obtained and available immediately, whereas transcripts are often available, if at all, with considerable delay.

In many applications, the ability to listen to recorded speech while at the same time having the visual channel available to perform other tasks is a significant benefit. For the blind, the auditory channel is probably the widest bandwidth channel available, and even normal recorded speech offers a "reading rate" that is typically 2 to 3 times that for Braille.

Finally, audio recordings have been used effectively in many learning situations, particularly, in the area of second-language learning.

The ability to modify the rate of speech is desirable for many reasons. The rate at which speech can be produced is constrained by physiological limitations to a maximum rate of about 110 to 180 wpm. However, the rate at which speech can be comprehended by most people is, typically, about 2 to 3 times this rate. Furthermore, without rate change, the rate of listening is completely paced by the recording and is not controllable by the listener. Consequently, the listener cannot scan or skip sections of the recording in the same manner as scanning printed text, nor can the listener slow down difficult-to-understand portions of the recording that might arise in the context of second language learning, or in listening to degraded speech. Clearly, the ability to modify the rate of (recorded) speech, while retaining its natural quality and intelligibility, would obviate these problems and have numerous applications. The modification of a speech signal such that the resulting signal differs from the original only by its perceived rate of articulation will, henceforth, be referred to as time-scale modification or rate change of speech.

## 1.2 HISTORICAL DEVELOPMENT OF THE PROBLEM

Although there are a number of presently available techniques for changing the rate of speech, they all introduce artifacts which degrade the quality of the processed speech. The most naive approach to time scaling speech is simply to play back recorded speech at a speed different from that at which it was recorded [Fletcher; Steinberg; Klumpp and Webster]. The problem here is obvious. Even with a small change in speed, spectral distortion is perceptible. As the difference between the record and playback speed is increased, the intelligibility deteriorates rapidly [Toong]. It is interesting to note that while most people are familiar with this effect, the time-scaled speech is generally described as changed in pitch. Although the pitch is, of course, changed, so is the spectral envelope of the speech. In fact, it is probably the frequency scaling of the spectral envelope and the corresponding shift in formant frequencies that contributes most to the degradation of the speech.

For the most part, nearly all algorithms for changing the rate of speech have been based on the Fairbanks technique [Fairbanks, et. al., 1954, 1959; Lee] or its refinement [Scott and Gerber; Huggins; Toong; Neuburg]. Basically, the Fairbanks scheme effects rate changes of speech by periodically repeating or discarding sections of the speech waveform. The duration of each section is chosen to be at least as long as one pitch period, but shorter than the length of a phoneme. This technique introduces discontinuities at the section boundaries which are perceived as "burbling" distortion and overall signal degradation.

The most popular refinement of the Fairbanks technique is pitch-synchronous implementation [Scott; Huggins; Toong]. Specifically, for portions of the speech that are voiced, the sections of speech that are repeated or discarded correspond to pitch periods. Although this scheme produces more intelligible speech than the basic asynchronous pitch-independent method, errors in pitch marking and voiced-unvoiced decisions introduce objectionable artifacts [Toong]. Moreover, since the ear is sensitive to these types of errors, and since pitch marking algorithms are generally sensitive to noise present in the speech, such algorithms would not be expected to be robust for noisy speech. Furthermore, even with no such detection errors, discontinuities may still be introduced [Toong].

Perhaps the most successful variant of the Fairbanks method is that recently proposed by [Neuburg]. This method uses a crude pitch detector, followed by an algorithm which repeats or discards sections of the speech equal in length to the average pitch period, then smooths together the edges of the sections that are retained. Because the method is not pitch synchronous, and, therefore, does not require pitch marking, it is more robust than pitch-synchronous implementations, yet much higher quality than pitch-independent methods.

Time-scale modification of speech based on classical vocoder methods [Schroeder; Flanagan] is an obvious approach. The speech would be represented by a set of time-varying parameters obtained as the output of the vocoder analyzer, the parameter tracks would be time scaled, and the rate-changed speech would then be generated by the vocoder synthesizer.

However, because the fundamental consideration in the formulation of a vocoder is bandwidth reduction, the vocoders currently available simply do not provide the high level of speech quality and naturalness we seek to attain. For example, a large class of vocoders require voiced-unvoiced decisions and pitch extraction. The resulting detection errors introduce artifacts to which the ear is particularly sensitive and which are not tolerable for our purpose.

Of the remaining classical vocoders, the only one that does not require voiced-unvoiced decisions and pitch extraction, yet is flexible enough to permit rate changes of speech is the phase vocoder [Flanagan and Golden; Schafer and Rabiner, 1973; Portnoff, 1976, 1977; Moorer]. The phase vocoder is a speech analysis / synthesis system based on short-time Fourier analysis and, unlike most vocoders, can be formulated to be an identity system in the absence of parameter modification. Furthermore, there is evidence that the ear is much less sensitive to errors in the short-time spectrum of an acoustic signal than to errors in the time-domain waveform [Callahan]. Unfortunately, because the theory of short-time Fourier analysis and its application to speech signals is not well understood, previous applications of the phase-vocoder to changing the rate of speech generally did not achieve the quality potentially attainable from this technique.

## 1.3 THE SCOPE OF THIS THESIS

This thesis deals with the problem of time-scale modification of speech based on short-time Fourier analysis. The objective is the development of a high-quality system for changing the rate of speech. The system must preserve such qualities as naturalness, intelligibility, and speaker dependent features. Furthermore, the system must not introduce such objectionable artifacts as "glitches," "burbles," or reverberation, often present in vocoded speech. Finally, the system should also be robust to noise, i.e., the performance of the system should not degrade severely if the source speech is corrupted by noise, as might occur in recordings of meetings or court-room proceedings.

The approach taken in this thesis is to formulate the problem of time-scale modification of speech in terms of three successive subproblems. The first of these is to appropriately model the speech signal. The second problem is to formulate a mathematical representation for the speech signal based on this model. This representation must have the property that time-scaling the parameters of the representation corresponds to the desired rate change of the speech signal. The third problem is to design and implement a high-quality analysis / synthesis system based on this representation. This system provides the means to manipulate the parameters of the speech model and should reduce to an identity system in the absence of any parameter modifications. In addition to these problems, various aspects of the theory of short-time Fourier analysis will be developed, in order to formulate the mathematics necessary to deal with the speech rate-change problem.

This thesis is divided into three parts. The first part, consisting of Chapters 2 and 3, develops the theory of short-time Fourier analysis for representing discrete-time signals and systems. The development in Chapter 2 is based on a continuous-frequency representation, whereas, the development in Chapter 3 is based on a discrete-frequency representation. The second part of the thesis, Chapter 4, develops a mathematical representation for the sampled speech signal based on the usual engineering model for speech production. This representation is used as the basis for defining rate-changed speech. The third, and final, part of the thesis, Chapters 5 and 6, applies the concepts of short-time Fourier analysis to speech to provide a mechanism for manipulating the speech parameters and thereby effect rate changes of the speech.

# CHAPTER 2

# TIME-FREQUENCY REPRESENTATION OF SIGNALS AND SYSTEMS

## 2.1 INTRODUCTION

The Fourier transform plays a fundamental role in the analysis of signals and linear time-invariant systems. The efficacy of the Fourier transform is a result of its providing a unique representation for signals in terms of the eigenfunctions of linear time-invariant systems, namely the complex exponentials. The essentials of this representation are summarized by the following well known results from the theory of linear time-invariant systems [Oppenheim and Schafer].

If $t(n)$ denotes the unit-sample (impulse) response of a linear time-invariant system, then the response, $y(n)$, of the system to the input $x(n)$ is given by the convolution sum

$$y(n) = \sum_{m=-\infty}^{\infty} t(n-m)\,x(m) = \sum_{m=-\infty}^{\infty} t(m)\,x(n-m) \quad . \qquad (2.1)$$

If $x(n)$ is the complex exponential $\exp[j\omega n]$ then eqn. (2.1) gives

$$y(n) = \sum_{m=-\infty}^{\infty} t(m)\exp[j\omega(n-m)]$$

$$= \Big(\sum_{m=-\infty}^{\infty} t(m)\exp[-j\omega m]\Big)\exp[j\omega n]$$

or,

$$y(n) = T(\omega)\exp[j\omega n] \qquad (2.2)$$

where $T(\omega)$, the frequency response of the system, is the Fourier transform of the unit-sample response given by

$$T(\omega) = \sum_{n=-\infty}^{\infty} t(n)\exp[-j\omega n] \quad . \qquad (2.3)$$

Suppose $x(n)$ is now a general signal that can be expressed as the Fourier integral

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)\exp[j\omega n]\,d\omega \qquad (2.4)$$

where $X(\omega)$ is the Fourier transform of $x(n)$ given by

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)\exp[-j\omega n] \quad . \qquad (2.5)$$

Eqns. (2.4) and (2.5) provide a unique correspondence between $x(n)$ and $X(\omega)$ and either one is an equally valid representation of the signal. The Fourier transform representation, however, is particularly convenient if the signal is to be processed by a linear time-invariant system because, the basis functions of the Fourier transform are the eigenfunctions of linear time-invariant systems. Specifically, since the Fourier integral (2.4) is, in essence, a linear combination of complex exponentials, and since the system $t(n)$ is linear, the response of $t(n)$ to the input (2.4) is

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} T(\omega) X(\omega) \exp[j\omega n] d\omega \qquad (2.6)$$

(assuming, of course, the sum (2.3) converges), and the product

$$Y(\omega) = T(\omega) X(\omega) \qquad (2.7)$$

is just the Fourier transform of the response, $y(n)$. Thus, the Fourier transform maps the convolution in the time domain to multiplication in the frequency domain. Furthermore, in addition to the Fourier transform being a powerful analytical technique, the property that its basis functions are the eigenfunctions of linear time-varying systems leads to a great deal of intuition, invaluable for solving signal-processing problems.

The Fourier-transform representation has several practical and conceptual limitations because it represents for each frequency $\omega$, the global (in time) characteristics of the signal. Consequently, the Fourier transform has the practical limitation that the entire signal must be known in order to obtain its transform. Moreover, the Fourier transform does not provide an adequate representation for linear time-varying systems, nor

does it always provide an intuitively meaningful representation for the output of such systems.

This chapter formulates a time-frequency representation for signals and linear time-varying systems that characterizes their local behavior in terms of complex exponentials. The representation of linear time-varying systems is based on the time-varying frequency response [Zadeh; Kailath; Gersho and DeClaris] and corresponds to a generalization of the frequency response (2.3) for linear time-invariant systems. The time-frequency representation for signals is based on the short-time Fourier transform [Gabor; Kharkevich; Weinstein; Callahan; Allen and Rabiner] which is a formal representation for the output of a filter-bank spectrum analyzer or, equivalently, the usual Fourier transform of the signal viewed through a sliding time window. The results derived here (and in Chapter 3) are of interest, in general, because they provide techniques applicable to a variety of signal-processing problems, and, specifically for the application in this thesis, because they provide both the mathematical framework for modelling rate changes of speech and the basis of an analysis / synthesis system capable of effecting high-quality rate changes of natural speech.

## 2.2 THE PARTIAL FOURIER TRANSFORM AND ITS INVERSE

The time-frequency representation developed in this thesis represents one-dimensional signals by two-dimensional signals. Because Fourier transforms with respect to one, or the other, or both, of the indices of such two-dimensional sequences frequently arise in this context, the definition and notation for such transforms will now be formalized. Let $f(n,m)$ denote a two-dimensional discrete-time sequence. The <u>complete</u> Fourier transform of $f(n,m)$, denoted $F(\psi,\omega)$, is defined as the (usual) two-dimensional Fourier transform

$$F(\psi,\omega) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} f(n,m) \exp[-j(\psi n+\omega m)] \tag{2.8}$$

with inverse transform

$$f(n,m) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} F(\psi,\omega) \exp[j(\psi n+\omega m)]\, d\psi d\omega \,. \tag{2.9}$$

The <u>partial</u> Fourier transform of $f(n,m)$ with respect to its first argument, denoted $F_1(\psi,m)$, is defined as the one-dimensional Fourier transform of $f(n,m)$ over $n$, that is,

$$F_1(\psi,m) = \sum_{n=-\infty}^{\infty} f(n,m) \exp[-j\psi n] \tag{2.10}$$

(where the subscript 1 is used to indicate that the first argument is the transform variable). Moreover, the inverse partial transform is given by

$$f(n,m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_1(\psi,m)\exp[j\psi n]\,d\psi \ . \qquad (2.11)$$

Similarly, the partial Fourier transform $F_2(n,\omega)$ is defined as the one-dimensional Fourier transform of $f(n,m)$ with respect to its second argument, i.e.,

$$F_2(n,\omega) = \sum_{m=-\infty}^{\infty} f(n,m)\exp[-j\omega m] \qquad (2.12)$$

and

$$f(n,m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_2(n,\omega)\exp[j\omega m]\,d\omega \ . \qquad (2.13)$$

Finally, the complete Fourier transform $F(\psi,\omega)$ can be obtained by successive partial Fourier transforms with respect to each of the independent arguments of $f(n,m)$, that is,

$$F(\psi,\omega) = \sum_{m=-\infty}^{\infty} F_1(\psi,m)\exp[-j\omega m] \qquad (2.14)$$

$$= \sum_{n=-\infty}^{\infty} F_2(n,\omega)\exp[-j\psi n] \ . \qquad (2.15)$$

## 2.3 THE TIME-VARYING FREQUENCY RESPONSE

The input-output behavior of a linear time-varying system can be characterized in the time-domain by a weighting pattern, or Green's function, $g(n,m)$ which represents the response of the system at time $n$ to a unit sample applied at time $m$. Equivalently, the same system can be described by a time-varying unit-sample response $t(n,m)$ defined as the response of the system at time $n$ to a unit sample applied $m$ samples earlier, i.e., at time $(n-m)$. Furthermore, the time-varying unit-sample response, $t(n,m)$, and the Green's function, $g(n,m)$, are related by

$$t(n,m) = g(n,n-m) \tag{2.16}$$

or, equivalently,

$$g(n,m) = t(n,n-m) \ . \tag{2.17}$$

If $y(n)$ is the response of a system to the input $x(n)$, then $y(n)$ is given by the superposition sum

$$y(n) = \sum_{m=-\infty}^{\infty} g(n,m)\,x(m) \tag{2.18}$$

or

$$y(n) = \sum_{m=-\infty}^{\infty} t(n,m)\,x(n-m) = \sum_{m=-\infty}^{\infty} t(n,n-m)\,x(m) \tag{2.19}$$

If the system represented by $g(n,m)$ is time-invariant, then $g(n,m)$ depends only on the difference $(n-m)$, corresponding to the number of samples between the application of the unit sample and the observation of the output; thus,

$$g(n,m) = g(n-m) \quad . \tag{2.20}$$

From the relation (2.16), the time-varying unit-sample response $t(n,m)$ for a linear time-invariant system becomes

$$\begin{aligned} t(n,m) &= g(n,n-m) = g(n-(n-m)) = g(m) \\ &= t(m) \end{aligned} \tag{2.21}$$

and corresponds to the ordinary unit-sample response of such a system. Conversely, if $t(n,m)$ is independent of $n$, then the system represented by $t(n,m)$ is time invariant. Moreover, if $t(n,m)$ is a "slowly-varying" function of $n$, then the system represented by $t(n,m)$ will be said to be slowly time varying. The notion of such a slowly time-varying system is, in general, imprecise and must be considered in the context of a particular set of assumptions about the system or the signals to be processed by the system. For example, a linear time-varying system may be said to be slowly time varying if the system can be regarded as stationary for the duration of its memory. A slowly time-varying system with this property will be referred to as a quasi-stationary system.

Because the time-varying unit-sample response is a characterization of a system relative to a sliding time frame, and because it is a slowly-varying function of $n$ for slowly-varying systems, the time-varying unit-sample response is more convenient than the weighting pattern in the context of short-time analysis. Therefore, the time-varying unit-sample response will be employed exclusively for the remainder of this thesis and referred to simply as the "unit-sample response."

If the input, $x(n)$, to a linear time-varying system with unit-sample response $t(n,m)$ is the complex exponential $\exp[j\omega n]$, then the resulting output is

$$y(n) = \sum_{m=-\infty}^{\infty} t(n,m)\,x(n-m)$$

$$= \sum_{m=-\infty}^{\infty} t(n,m)\exp[j\omega(n-m)]$$

$$= \sum_{m=-\infty}^{\infty} t(n,m)\exp[-j\omega m]\exp[j\omega n]$$

or

$$y(n) = T_2(n,\omega)\exp[j\omega n] \qquad (2.22)$$

where

$$T_2(n,\omega) = \sum_{m=-\infty}^{\infty} t(n,m)\exp[-j\omega m] \; . \qquad (2.23)$$

$T_2(n,\omega)$, the partial Fourier transform of $t(n,m)$ with respect to $m$, is interpreted according to eqn. (2.22) as the <u>time-varying frequency response</u> of the system with unit-sample response $t(n,m)$. For simplicity, $T_2(n,\omega)$ will often be referred to, simply, as the "frequency response" of $t(n,m)$.

If $X(\omega)$ is the Fourier transform of an arbitrary input, $x(n)$, then the response, $y(n)$, of $t(n,m)$ can be expressed as the inverse partial Fourier transform of the product of $X(\omega)$ with $T_2(n,\omega)$, that is,

$$y(n) = \sum_{m=-\infty}^{\infty} t(n,m)\,x(n-m)$$

$$= \sum_{m=-\infty}^{\infty} t(n,m) \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) \exp[j\omega(n-m)]\, d\omega \right)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left( \sum_{m=-\infty}^{\infty} t(n,m) \exp[-j\omega m] \right) X(\omega) \exp[j\omega n]\, d\omega$$

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} T_2(n,\omega) X(\omega) \exp[j\omega n]\, d\omega \qquad (2.24)$$

and eqn. (2.24) is a generalization of eqn. (2.6) for linear time-invariant systems. In contrast to the case of linear time-invariant systems, however, it is not generally true that the time-varying frequency response of the cascade combination of linear time-varying systems is equal to the product of the corresponding individual time-varying frequency responses. In fact, there exists no such scalar-valued function with this property because the input-output behavior of such a cascade combination of systems depends on the order in which the systems are cascaded.

For the case of a linear time-invariant system, $t(n,m)$ and, hence, $T_2(n,\omega)$ are independent of n. Thus, $t(n,m) = t(m)$ and $T_2(n,\omega) = T(\omega)$ are the ordinary unit-sample response and frequency response for such a system.

## 2.4 SHORT-TIME FOURIER ANALYSIS AND SYNTHESIS

The usual short-time Fourier transform representation for a discrete-time signal $x(n)$ is given by the pair of equations [Weinstein]

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X_2(n,\omega)\exp[j\omega n]\,d\omega \qquad (2.25)$$

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\,x(m)\exp[-j\omega m] \qquad (2.26)$$

where, using the notation of Section 2.2, $X_2(n,\omega)$ denotes the short-time Fourier transform of $x(n)$. $h(n)$ is referred to as the analysis window and is generally chosen to have the property that it is, in some sense, narrow in time, or frequency, or both, and is normalized such that $h(0) = 1$. Eqn. (2.25) is similar in form to the ordinary Fourier synthesis relation (2.4) except that $X_2(n,\omega)$ is now a function of the time index $n$ and represents only the local behavior of $x(m)$ as viewed through the sliding window $h(n-m)$. Referring to Figure 2.1, $X_2(n,\omega)$ can be interpreted for each value of $n$ as the partial Fourier transform, with respect to $m$, of the "short-time function"

$$x(n,m) = h(n-m)\,x(m)\ . \qquad (2.27)$$

Equivalently, by considering eqn. (2.26) as the convolution

$$X_2(n,\omega) = h(n) *_n x(n)\exp[-j\omega n]\ , \qquad (2.28)$$

where $*_n$ denotes the convolution operator with respect to $n$, $X_2(n,\omega)$ can be interpreted as the output of a linear time-invariant filter $h(n)$, excited by the demodulated (frequency-shifted) signal $x(n)\exp[-j\omega n]$, as shown in Figure 2.2. For this reason, $h(n)$ is also referred to as the analysis filter. Because $X_2(n,\omega)$ is a function of the continuous variable $\omega$ for every value of $n$, the short-time Fourier transform contains redundant

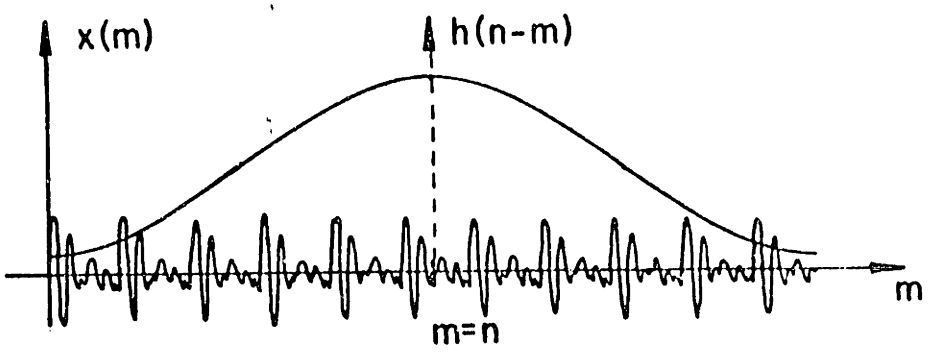



Figure 2.1(a)

Analysis Window h(n) Shifted and Superimposed on Data x(m)


Figure 2.1(b)

Short-Time Sequence $x(n,m) = h(n-m)x(m)$ for a Particular Value of n

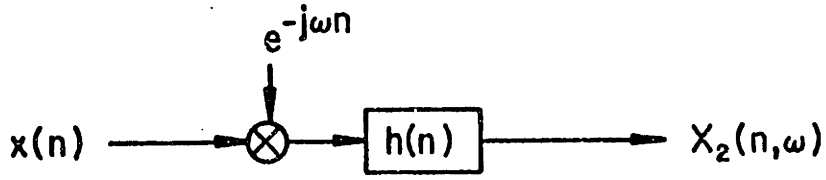



Figure 2.2

Short-Time Fourier Transform as Output of Demodulator Followed by Analysis Filter

information about the signal, depending upon the particular analysis window used in eqn. (2.26). Furthermore, eqn. (2.26) imposes a structure on $X_2(n,\omega)$ so that not all functions of n and $\omega$ are valid short-time Fourier transforms.

To illustrate the structure imposed on $X_2(n,\omega)$ by eqn. (2.26), observe that the inverse Fourier transform of $X_2(n,\omega)$ with respect to $\omega$ is the short-time function $x(n,m)$, which factors as the product of the signal multiplied by the shifted window, i.e.,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} X_2(n,\omega)\exp[j\omega m]\,d\omega = x(n,m)$$

$$= h(n-m)x(m) \quad . \tag{2.29}$$

Thus, not only can the signal $x(n)$ be recovered from the short-time Fourier transform, by evaluating eqn. (2.29) for $n = m$, but the analysis window, $h(n)$, can also be recovered, to within the multiplicative constant $x(0)$, by evaluating eqn. (2.29) for $m = 0$. In addition to the convolutional structure of eqn. (2.26), $X_2(n,\omega)$ also exhibits a convolutional structure when expressed in terms of the Fourier transforms of $h(n)$ and $x(n)$. Replacing $h(n-m)$ and $x(m)$ in eqn. (2.26) by their Fourier integral representations and simplifying gives $X_2(n,\omega)$ as the frequency domain convolution

$$X_2(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega+\psi)\,H(\psi)\exp[j\psi n]\,d\psi \tag{2.30a}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\psi)\,H(\psi-\omega)\exp[j(\psi-\omega)n]\,d\psi \tag{2.30b}$$

$$= \frac{1}{2\pi}\,X(\omega) \;*_{\omega}\; H(\omega)\exp[j\omega n]\ . \tag{2.30c}$$

Furthermore, the partial Fourier transform of $X_2(n,\omega)$ with respect to n is obtained by inspection of eqn. (2.30a) as

$$X(\psi,\omega) = H(\psi)\,X(\omega+\psi) \tag{2.31}$$

and also factors as the product of a function which depends only on the window and a function that depends only on the signal. Finally, $X(\psi,\omega)$ is recognized as the two-dimensional Fourier transform of the short-time function $x(n,m)$, that is,

$$X(\psi,\omega) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} h(n-m)\,x(m)\exp[-j(\psi n+\omega m)]$$

$$= \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} x(n,m)\exp[-j(\psi n+\omega m)]\ . \tag{2.32}$$

As a result of the mathematical structure of $X_2(n,\omega)$, eqn. (2.25) is not the only means for synthesizing $x(n)$ from $X_2(n,\omega)$. Eqn. (2.25) corresponds to inverse transforming $X_2(n,\omega)$ with respect to $\omega$ to obtain the short-time function (2.27), which is then evaluated at $m = n$ to give

$$x(n,m)\Big|_{m=n} = h(0)x(n) = x(n) \qquad \text{for } h(0) = 1. \tag{2.33}$$

Alternatively, $x(m)$ could be obtained by again inverse transforming $X_2(n,\omega)$ to get $x(n,m) = h(n-m)x(m)$, but now, fixing $n = n_0$ and dividing by the shifted window, $h(n_0-m)$, i.e.,

$$x(m) = (1/2\pi h(n_0-m)) \int_{-\pi}^{\pi} X_2(n_0,\omega) \exp[j\omega m]\, d\omega \tag{2.34}$$

$$= x(n_0,m)/h(n_0-m) = [\, h(n_0-m)x(m)\,]/h(n_0-m)$$

$$= x(m) \ .$$

Clearly, for the particular value $n_0$, eqn. (2.34) is useful only for obtaining values of $x(m)$ where $h(n_0-m) \neq 0$. Another method of short-time Fourier synthesis [Parsons, Allen] can be derived by evaluating $X(\psi,\omega)$, given by eqn. (2.31), for $\psi = 0$ to obtain

$$X(\omega) = X(0,\omega) \ / \ H(0)$$

$$= \frac{1}{H(0)} \sum_{n=-\infty}^{\infty} X_2(n,\omega) \ . \tag{2.35}$$

Since $x(n)$ is the inverse Fourier transform of $X(\omega)$:

$$x(n) = (1/2\pi H(0)) \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} X_2(r,\omega) \exp[j\omega n]\, d\omega \ . \tag{2.36}$$

From the definition (2.26) of the short-time Fourier transform, its value at a particular time sample $n = n_0$ represents information not only about $x(n_0)$, but about all values of of $x(n)$ "viewed" through the sliding (time) window $h(n_0-n)$. Similarly, from eqn. (2.30b), the short-time Fourier

transform evaluated at a particular frequency $\omega = \omega_0$, contains information about all values of $X(\omega)$ viewed through the sliding (frequency) window $H(\omega-\omega_0)$. Thus, the values of $X_2(n,\omega)$ are locally correlated in time and frequency. The synthesis formula (2.25) corresponds to the inverse partial Fourier transform of $X_2(n,\omega)$ with respect to $\omega$, evaluated for $n = m$ : for a particular value $n = n_0$, $x(n_0)$ is computed solely from the values of $X_2(n_0,\omega)$, ignoring the local correlation of the values of $X_2(n,\omega)$ in time . The synthesis formula (2.34) corresponds to the inverse Fourier transform of $X_2(n,\omega)$ evaluated at $n = n_0$, so that for all values of n, the values of $x(n)$ are computed from $X_2(n,\omega)$ evaluated only at the particular sample $n = n_0$. The synthesis formula (2.36) corresponds to the inverse Fourier transform of

$$X(0,\omega) = \sum_{n=-\infty}^{\infty} X_2(n,\omega) \ .$$

Although this synthesis procedure utilizes information from adjacent time samples of $X_2(n,\omega)$, it simply sums over all n, giving equal weight to each value.

All of these synthesis formulae can be viewed in a more general framework by exploiting the local correlation of the values of $X_2(n,\omega)$. Introducing a "synthesis window," denoted $F_1(\omega,n)$, and formulating a new synthesis equation by replacing $X_2(n,\omega)$ in the conventional synthesis formula (2.25) by the moving average

$$X_2'(n,\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} F_1(\omega-\varphi, n-r) X_2(r,\varphi) \, d\varphi \qquad (2.37)$$

gives

$$x(n) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} F_1(\omega-\varphi, n-r)\, X_2(r,\varphi) \exp[j\omega n]\, d\varphi d\omega \qquad (2.38)$$

where $F_1(\omega, n)$ depends on the analysis window and remains to be determined. Eqn. (2.38) can be simplified by performing the integration with respect to $\omega$ to obtain

$$x(n) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} \left( \int_{-\pi}^{\pi} F_1(\omega-\varphi, n-r) \exp[j(\omega-\varphi)n]\, d\omega \right) \times X_2(r,\varphi) \exp[j\varphi n]\, d\varphi$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n, n-r)\, X_2(r,\varphi) \exp[j\varphi n]\, d\varphi$$

or, replacing $\varphi$ with $\omega$,

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n, n-r)\, X_2(r,\omega) \exp[j\omega n]\, d\omega \,. \qquad (2.39)$$

The three synthesis formulae (2.25), (2.34), and (2.36) now become the special cases:

$$f(n, m) = \delta[m] \qquad (2.40a)$$

$$f(n, m) = \delta[n-m-n_0]/h(-m) \qquad (2.40b)$$

and

$$f(n, m) = 1 \qquad (2.40c)$$

respectively, where $\delta[n]$ denotes the unit sample.

To derive the general relationship between $f(n,m)$ and $h(n)$ so that eqn. (2.39) synthesizes $x(n)$ from $X_2(n,\omega)$, interchange the order of integration and summation in eqn. (2.39) and recognize the integral as the inverse partial Fourier transform of $X_2(n,\omega)$, which is just the short-time function $x(r,n) = h(r-n)x(n)$. Eqn. (2.39), therefore, reduces to

$$x(n) = \sum_{r=-\infty}^{\infty} f(n,n-r)\,h(r-n)\,x(n) \tag{2.41}$$

$$= \Big(\sum_{m=-\infty}^{\infty} f(n,-m)\,h(m)\Big)x(n)$$

$$= x(n)$$

if and only if

$$\sum_{m=-\infty}^{\infty} f(n,-m)\,h(m) = 1\ , \tag{2.42}$$

or equivalently, if and only if

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F_2(n,\omega)\,H(\omega)\,d\omega = 1\ . \tag{2.43}$$

The synthesis procedure, implied by eqn. (2.39), has two interpretations depending on the interpretation of the short-time Fourier transform. If $X_2(n,\omega)$ is interpreted as a set (indexed by $\omega$) of sequences in $n$, then the synthesis procedure corresponds to filtering $X_2(n,\omega)$ with the linear time-varying filter $f(n,m)$ to obtain the set (also indexed by $\omega$) of sequences in $n$

$$Z_2(n,\omega) = \sum_{r=-\infty}^{\infty} f(n,n-r)\,X_2(r,\omega). \tag{2.44}$$

$x(n)$ is obtained by modulating $Z_2(n,\omega)$ by $\exp[j\omega n]$ for every value of $\omega$ and integrating over $\omega$ as illustrated in Figure 2.3. Consequently, $f(n,m)$ will also be referred to as the "synthesis filter." Alternatively, if $X_2(r,\omega)$ is interpreted as a set (indexed by r) of Fourier transforms, then eqn. (2.39) can be viewed as inverse Fourier transforming $X_2(r,\omega)$ to obtain the set (also indexed by r) of short-time sequences in n

$$\begin{aligned} \xi(r,n) &= f(n,n-r)\,x(r,n) \\ &= f(n,n-r)\,h(r-n)\,x(n). \end{aligned} \tag{2.45}$$

where $\xi(r,n)$ corresponds to $x(n)$ weighted by the time-varying shifted window

$$f(n,-m)\,h(m)\Big|_{m = r-n}.$$

For each value of n, $x(n)$ is obtained by projecting (summing) $\xi(r,n)$ in r, according to eqn. (2.41), as illustrated in Figure 2.4.

A more general theory of short-time Fourier analysis and synthesis can be formulated by also allowing the analysis window to vary as a function of time. Such a formulation is appropriate for adaptive-processing schemes when the particular choice of analysis window depends on the data to be analyzed [Patisual and Hammett]. In addition, both the analysis and synthesis windows can be allowed to be functions of frequency, for such techniques as constant-Q analysis [Gambardella, 1971; Youngberg and Boll]. These formulations are more general than required for the application in this thesis and will not be pursued here. Furthermore, for the remainder

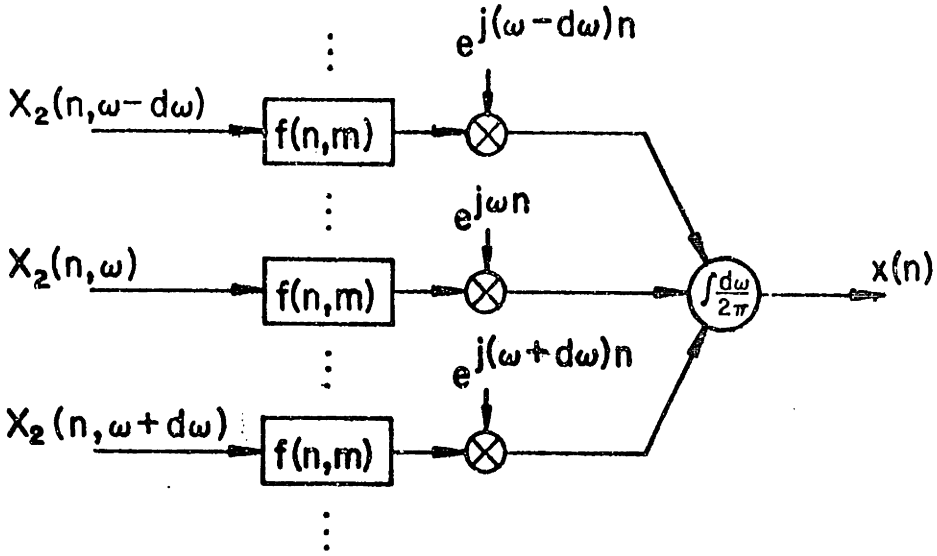



Figure 2.3

Synthesis of Time Sequence as Combination of Filterbank Outputs

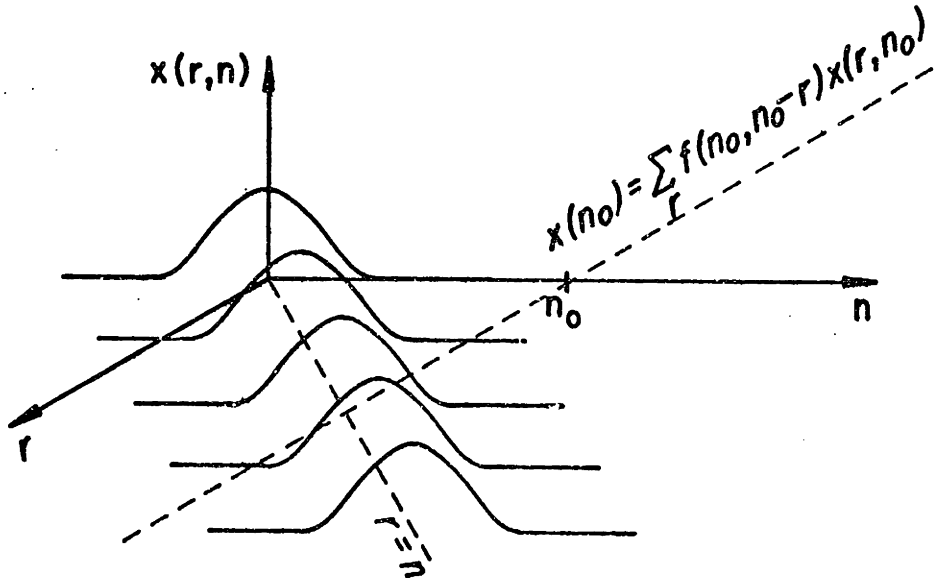



Figure 2.4

Synthesis of Time Sequence as Weighted Projection of Short-Time Sequence

of this thesis, only time-invariant synthesis filters will be considered. Thus, the synthesis formula (2.39) reduces to

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n-r) X_2(r,\omega) \exp[j\omega n] d\omega \qquad (2.46)$$

where the analysis and synthesis filters satisfy the condition

$$\sum_{m=-\infty}^{\infty} f(-m) h(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega) H(\omega) d\omega = 1 . \qquad (2.47)$$

This mild condition is the time-invariant form of the conditions (2.42) and (2.43) and can be satisfied, simply, by appropriately scaling the analysis and synthesis filters, provided that the summation and integral in eqn. (2.47) do not vanish.

## 2.5 SHORT-TIME FOURIER ANALYSIS OF NARROW-BAND SIGNALS

In many applications, signals arise that occupy a narrow band of frequencies. Within the context of short-time Fourier analysis, the term "narrow-band signal" will mean that the bandwidth of the signal is narrow compared with the bandwidth of the analysis filter. Let $x(n)$ represent such a signal occupying a narrow band of frequencies centered about $\omega_0$. Its Fourier transform, $X(\omega)$, is illustrated in Figure 2.5a, and its short-time Fourier transform, $X_2(n,\omega)$, is given, in terms of $X(\omega)$, by eqn. (2.30b), as

$$X_2(n,\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\psi) H(\psi-\omega) \exp[j(\psi-\omega)n] d\psi . \qquad (2.48)$$

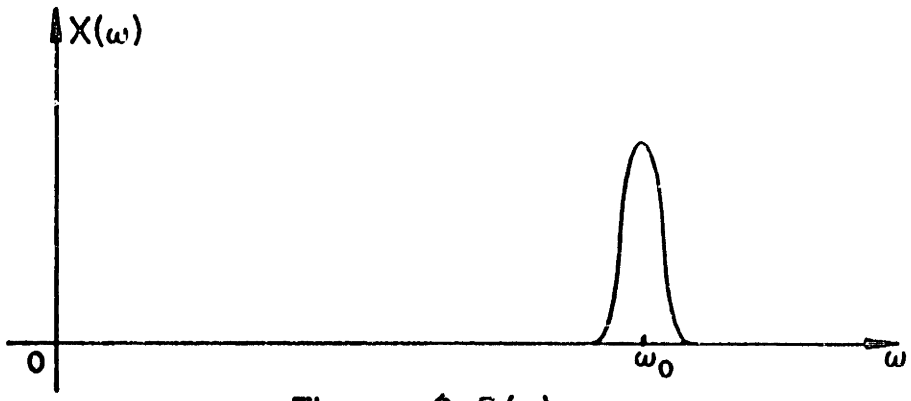



Figure 2.5(a)

Fourier Transform of a Narrow-Band Signal

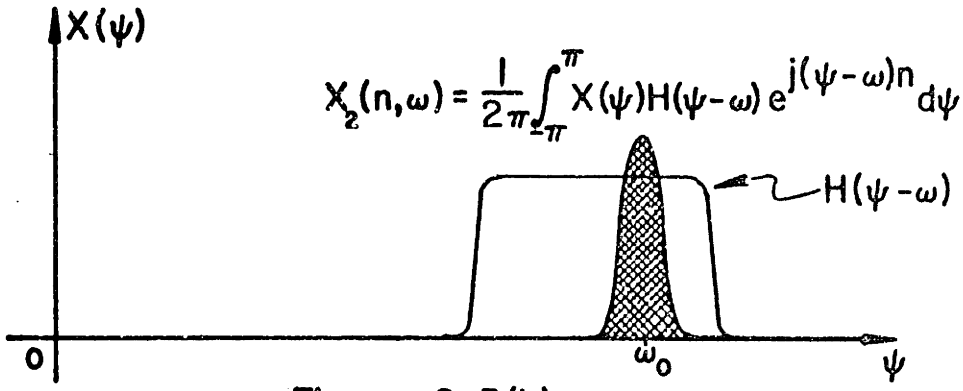



Figure 2.5(b)

Short-Time Fourier Transform as a Convolution Integral with $\psi$ as Variable of Integration

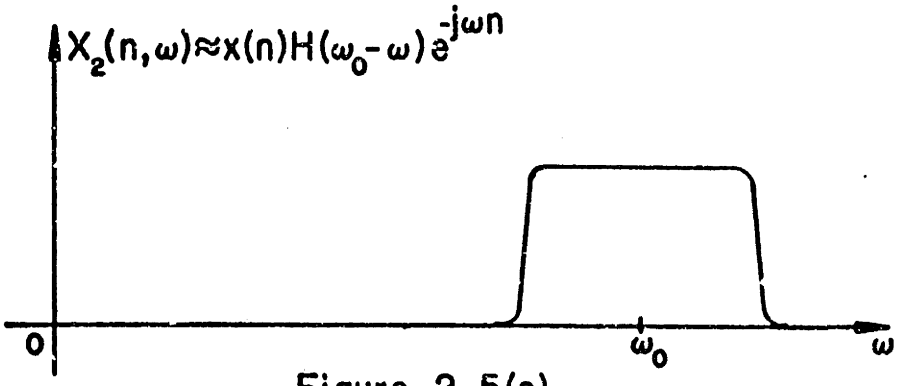



Figure 2.5(c)

Short-Time Fourier Transform of a Narrow-Band Signal

Interpreting eqn. (2.48) as a convolution integral and referring to Figure 2.5b, $X_2(n,\omega)$ can be evaluated by approximating $H(\psi-\omega)$ as constant in the region overlapping $X(\psi)$. Thus, the product $X(\psi)H(\psi-\omega)$ is replaced by $X(\psi)H(\omega_0-\omega)$, and eqn. (2.48) becomes

$$X_2(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\psi)H(\omega_0-\omega)\exp[j(\psi-\omega)n]d\psi$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\psi)\exp[j\psi n]d\psi \; H(\omega_0-\omega)\exp[-j\omega n].$$

Integrating over $\psi$ gives the desired result

$$X_2(n,\omega) = x(n)H(\omega_0-\omega)\exp[-j\omega n] . \qquad (2.49)$$

Figure 2.5c illustrates $X_2(n,\omega)$, for a fixed value of n, as the image of $H(-\omega)$ shifted by $\omega_0$ and weighted by $x(n)\exp[-j\omega n]$.

## 2.6 LINEAR FILTERING BASED ON SHORT-TIME FOURIER ANALYSIS

In view of the property of the Fourier transform that maps convolution in one domain to multiplication in the other domain and the generalization of this property to eqn. (2.24) for linear time-varying systems, it is natural to ask whether an analogous property exists for short-time Fourier transforms. Such a property does, indeed, exist, but applies only to a restricted class of linear time-varying systems determined by the filters used in the short-time Fourier analysis and synthesis. Specifically, this property is the following. Let $X_2(n,\omega)$ denote the short-time Fourier transform of $x(n)$ and let $T_2(n,\omega)$ denote the frequency response of an

arbitrary linear time-varying system $t(n,m)$. Further, define the modified short-time Fourier transform $Y_2(n,\omega)$ as the product

$$Y_2(n,\omega) = T_2(n,\omega)X_2(n,\omega) \quad , \tag{2.50}$$

noting that, in general, $Y_2(n,\omega)$ is not a valid short-time Fourier transform, in the sense that it cannot be expressed in the form of eqn. (2.26). If $y(n)$ is synthesized from $Y_2(n,\omega)$ according to the formula

$$y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty} f(n-r)Y_2(r,\omega)\exp[j\omega n]\,d\omega \tag{2.51}$$

then

$$y(n) = \sum_{m=-\infty}^{\infty} \bar{t}(n,m)x(n-m) \tag{2.52}$$

and corresponds to the response, to $x(n)$, of the linear time-varying system with unit-sample response $\bar{t}(n,m)$, given by

$$\begin{aligned}\bar{t}(n,m) &= \sum_{r=-\infty}^{\infty} f(r)h(m-r)t(n-r,m) \\ &= \{f(n)h(m-n)\} \circledast_n t(n,m)\end{aligned} \tag{2.53}$$

Furthermore, the modified system given by eqn. (2.53) is conveniently characterized as the product of the partial Fourier transforms:

$$\bar{T}_1(\psi,m) = F_2(m,\psi)T_1(\psi,m) \tag{2.54}$$

where

$$F_2(m,\psi) = \sum_{r=-\infty}^{\infty} f(r)h(m-r)\exp[-j\psi r]$$

denotes the short-time Fourier transform of the synthesis filter f(n) (not to be confused with the partial Fourier transform of the now discarded time-varying synthesis filter, f(n,m)).

The proof of eqns. (2.52) and (2.53) follows from substituting eqn. (2.50) into eqn. (2.51) to obtain

$$y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty} f(n-r)\,T_2(r,\omega)\,X_2(r,\omega)\exp[j\omega n]\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty} f(n-r)\,T_2(r,\omega) \times \sum_{m=-\infty}^{\infty} h(r-m)\,x(m)\exp[-j\omega m]\exp[j\omega n]\,d\omega$$

$$= \sum_{m=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} f(n-r)\,h(r-m) \times \frac{1}{2\pi}\int_{-\pi}^{\pi} T_2(r,\omega)\exp[j\omega(n-m)]\,d\omega\; x(m)$$

$$= \sum_{m=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} f(n-r)\,h(r-m)\,t(r,n-m)\,x(m)\ .$$

Letting $m' = n-m$ and $r' = n-r$ gives

$$y(n) = \sum_{m'=-\infty}^{\infty}\sum_{r'=-\infty}^{\infty} f(r')\,h(m'-r')\,t(n-r',m')\,x(n-m')$$

from which eqns. (2.52) and (2.53) follow.

Thus, for a particular pair of analysis and synthesis filters, linear filtering implemented as the product of an time-varying frequency response and the short-time Fourier transform of the input sequence, is restricted

to the class of filters with unit-sample responses of the form $\bar{t}(n,m)$ specified by eqn. (2.53), with $t(n,m)$ arbitrary. The significance of this condition, a limitation on the simultaneous time and frequency variation of the filter $\bar{t}(n,m)$, determined by the short-time Fourier analysis and synthesis filters, is illustrated by considering three special cases.

First, a linear time-varying filter $\bar{t}(n,m)$ with arbitrary time-variation in $n$ can be implemented for a given analysis window, $h(n)$, by designing the synthesis window $f(n)$ to be of much shorter duration than $h(n)$. In this case, $h(n)$ is approximately constant over the duration of $f(n)$ and $t(n,m)$ becomes

$$\bar{t}(n,m) = \sum_{r=-\infty}^{\infty} f(r)\,h(m-r)\,t(n-r,m)$$

$$\approx \sum_{r=-\infty}^{\infty} f(r)\,h(m)\,t(n-r,m)$$

$$= \sum_{r=-\infty}^{\infty} f(n-r)\,t(r,m)\,h(m)$$

$$\bar{t}(n,m) = f(n) *_n t(n,m)\,h(m) \,. \qquad (2.55)$$

Thus, the unit-sample response $t(n,m)$ is windowed by $h(m)$ in the $m$ direction and smoothed by $f(n)$ in the $n$ direction. Equivalently, the time-varying frequency response $T_2(n,\omega)$ is smoothed by $H(\omega)$ in $\omega$ and by $f(n)$ in $n$. In the limit, $f(n)$ can be chosen as $f(n) = \delta[n]$ so that $\bar{t}(n,m)$ becomes (exactly)

$$\bar{t}(n,m) = t(n,m)\,h(m) \,. \qquad (2.56)$$

The second case is complementary to the first. Namely, a linear time-varying filter with arbitrary frequency variation in $\omega$ can be implemented by designing the synthesis window $f(n)$ to be of much greater duration than $h(n)$. Here, $f(n)$ is approximately constant over the duration of $h(n)$ so that $\bar{t}(n,m)$ becomes

$$\bar{t}(n,m) = \sum_{r=-\infty}^{\infty} f(r)\,h(m-r)\,t(n-r,m)$$

$$\approx \sum_{r=-\infty}^{\infty} f(m)\,h(m-r)\,t(n-r,m)$$

$$= \sum_{r=-\infty}^{\infty} h(m-n-r)\,t(r,m)\,f(m)$$

$$\bar{t}(n,m) = h(r) \ast_r t(r,m)\Big|_{r=m-n} f(m) \,. \qquad (2.57)$$

In this case $t(n,m)$ is windowed by $f(m)$ in $m$ and smoothed by $h(n)$ in $n$, but an additional "smearing" is introduced because the convolution of eqn. (2.57) is evaluated for $(m-n)$ rather than $n$. In the limit, $f(n)$ can be chosen as unity and $\bar{t}(n,m)$ becomes

$$\bar{t}(n,m) = h(r) \ast_r t(r,m)\Big|_{r=m-n} \,. \qquad (2.58)$$

The third special case is the implementation of a slowly time-varying filter. If $t(n,m)$ can be approximated as stationary over the duration of the synthesis window, then $\bar{t}(n,m)$ becomes

$$\bar{t}(n,m) = \sum_{r=-\infty}^{\infty} f(r)\,h(m-r)\,t(n-r,m)$$

$$\approx \sum_{r=-\infty}^{\infty} f(r)\,h(m-r)\,t(n,m)$$

$$\bar{t}(n,m) = \{f(m) *_m h(m)\} t(n,m) \qquad (2.59)$$

Thus, $\bar{t}(n,m)$ is just the product of $t(n,m)$ with the effective window $f(m) * h(m)$. Alternatively, eqn. (2.59) is also valid if $t(n,m)$ is slowly-varying in the sense that the bandwidth of $T_1(\psi,m)$ is narrow compared with the bandwidth of $F(\psi)$.

In practice, the limit on simultaneous time and frequency variation is generally not a serious restriction. In fact, it can often be exploited. A common application of short-time Fourier analysis is for adaptive filtering. Here, a signal is filtered by a time-varying system, the characteristics of which depend on the local characteristics of the input signal. By using the formulation leading to eqn. (2.56), filter design by windowing [Oppenheim and Schafer] can be accomplished automatically. Furthermore, by using the formulation leading to eqn. (2.55), linear smoothing of the time variation of the time-varying frequency response can also be introduced.

# CHAPTER 3

# DISCRETE TIME-FREQUENCY REPRESENTATION OF SIGNALS AND SYSTEMS

## 3.1 INTRODUCTION

In order to realize a signal processing algorithm based on the time-frequency representation discussed in Chapter 2, on a digital processor, the short-time Fourier transform and time-varying frequency response must be represented by a finite number of frequency samples. Moreover, to make the amount of computation tractable, the short-time Fourier transform must be decimated in time (down sampled) as well.

This chapter extends the results of Chapter 2 to a sampled-transform representation based on the discrete (sampled) short-time Fourier transform and the discrete time-varying frequency response. The development focuses on three problems. The first is the representation of a sequence in terms of samples of its short-time Fourier transform and the resynthesis of the

original sequence without distortion. Of special interest is the problem of formulating such a representation with no redundancy, i.e., so that there is, on the average, one sample of the transform representation for each sample of the original signal. The second problem is the efficient implementation of the discrete short-time Fourier analysis and synthesis formulae based on the fast Fourier transform (FFT) algorithm. Because the short-time Fourier analysis and synthesis formulae do not have the form of discrete Fourier transforms (DFT) they cannot be computed directly with the FFT algorithm. The third, and final, problem considered in this chapter is the implementation of linear time-varying filtering as the product of the discrete short-time Fourier transform of the signal to be filtered multiplied by samples of the time-varying frequency response of the filter. For a class of linear time-varying filters, determined by the short-time Fourier analysis and synthesis filters, such an implementation is possible, and for linear time-invariant filters, this implementation reduces to the conventional overlap-save or overlap-add technique of fast convolution, depending on the particular choice of the analysis and synthesis filters.

## 3.2 DISCRETE SHORT-TIME FOURIER ANALYSIS AND SYNTHESIS

The short-time Fourier transform $X_2(n,\omega)$ represents the sequence $x(n)$ by a function of the continuous variable $\omega$ for each value of the index n and contains redundant information about the signal and the analysis window. This section considers the problem of representing a sequence by samples of its short-time Fourier transform with the result that for proper

choices of sampling rates in both time and frequency and for certain choices of analysis and synthesis filters, the short-time Fourier transform $X_2(n,\omega)$ of $x(n)$ can be sampled using, on the average, one sample per value of $x(n)$.

Define the discrete short-time Fourier transform of $x(n)$ as

$$X_2(sR, k\Omega_0) = \sum_{m=-\infty}^{\infty} h(sR-m)\,x(m)\exp[-j\Omega_0 km] \qquad (3.1)$$

corresponding to samples of the short-time Fourier transform specified every R samples in time and $\Omega_0 = 2\pi/M$ radians in frequency. For certain choices of the sampling parameters, R and M, and the filters, $h(n)$ and $f(n)$, $x(n)$ can be recovered by means of the synthesis formula

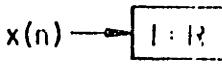
$$x(n) = \frac{1}{M}\sum_{k=0}^{M-1}\sum_{s=-\infty}^{\infty} f(n-sR)\,X_2(sR, k\Omega_0)\exp[j\Omega_0 kn]\ . \qquad (3.2)$$

The pair of equations (3.1) and (3.2) describe the M-channel filter bank analysis / synthesis system depicted in Figure 3.1. $X_2(sR, k\Omega_0)$ is the output of the analysis filter, $h(n)$, in the kth channel, uniformly sampled every R samples. The synthesized signal is generated by interpolating (time expanding and filtering) each of the M channels with $f(n)$, modulating by $\exp[j\Omega_0 kn]$, and summing over all the channels.

Referring to Figure 3.1, the identical combination of the analysis filter, $h(n)$, followed by an R:1 compressor, and a 1:R interpolator, (1:R expander followed by the synthesis filter, $f(n)$), appears in each of the M channels of the filterbank. This combination can be replaced in each channel, as illustrated in Figure 3.2, by the equivalent linear time-varying system $w(n,m)$ given by

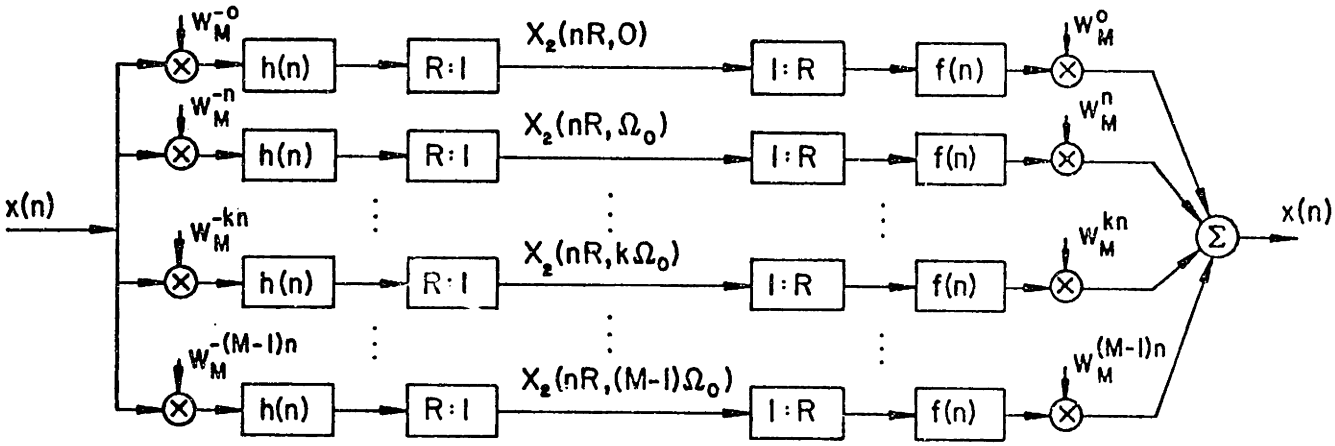

$y(n) = x(nR)$

Figure 3.1(a)
R:1 Compressor




$$y(n) = \begin{cases} x(n/R), & n = 0, \pm R, \pm 2R, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Figure 3.1(b)
1:R Expander




$W_M = e^{j\Omega_0}$

$\Omega_0 = 2\pi/M$

Figure 3.1(c)

Filterbank Analogue for Discrete Short-Time Fourier Analysis / Synthesis

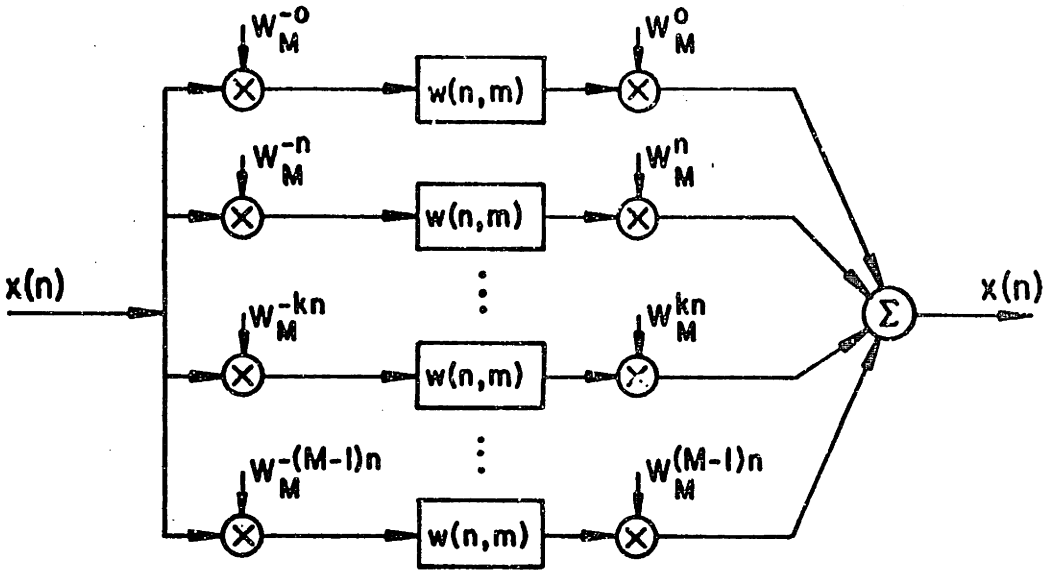



Figure 3.2(a)
Filterbank Equivalent to Figure 3.1(c)

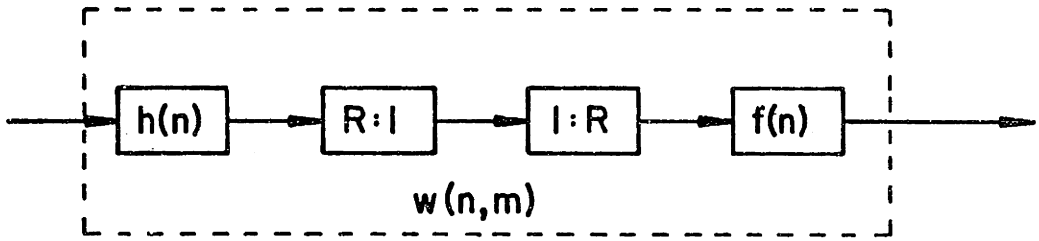



Figure 3.2(b)
Equivalent Linear Time-Varying Filter
for each Channel

$$w(n,m) = \sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR-n+m) \quad . \tag{3.3}$$

Recall that $w(n,m)$ denotes the response of the system at sample n to a unit sample applied at time $(n-m)$. For each value of n, eqn. (3.3) corresponds to uniformly sampling $h(n)$ every R samples, starting at sample $(n-m)$, then interpolating by 1:R with $f(n)$. The time variation of $w(n,m)$ results because decimation / interpolation is not a time-invariant operation. The condition such that $x(n)$ can be recovered exactly from $X_2(sR, k\Omega_0)$ by eqn. (3.2) is

$$w(n,pM) = \delta[p] \qquad \text{for all } n. \tag{3.4}$$

This result follows from substituting the definition (3.1) of $X_2(sR, k\Omega_0)$ into eqn. (3.2) to obtain

$$x(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR) \left( \sum_{m=-\infty}^{\infty} h(sR-m)\, x(m) \exp[-j\Omega_0 km] \right) \exp[j\Omega_0 kn]$$

$$= \sum_{s=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} f(n-sR)\, h(sR-m)\, x(m) \left( \frac{1}{M} \sum_{k=0}^{M-1} \exp[j\Omega_0 k(n-m)] \right)$$

or,

$$x(n) = \sum_{p=-\infty}^{\infty} f(n-sR)\, h(sR-n+pM)\, x(n-pM)$$

if and only if

$$\sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR-n+pM) = \delta[p] \qquad \text{for all } n. \tag{3.5}$$

Using the definition (3.3) for $w(n,m)$ yields the desired condition (3.4).

The representation of a sequence by its sampled short-time Fourier transform with no redundancy results when the decimation ratio, R, is equal to the number of frequency samples, M. In this case, eqns. (3.3) and (3.4) become

$$\sum_{s=-\infty}^{\infty} f(n-sM)\, h((p+s)M-n) = \delta[p] \qquad \text{for all } n. \tag{3.6}$$

Because the left-hand side of eqn. (3.6) is periodic in n with period M, the problem of determining the synthesis filter f(n), given a particular analysis window, h(n), reduces to M independent inverse filtering problems, one for each value of n in the range $0 \leq n < M$.

An alternate formulation of the condition for the exact resynthesis of x(n) according to eqn. (3.2), can be obtained by taking the partial Fourier transform of eqn. (3.5) with respect to n, giving

$$F_2(pM,\omega)\left\{\frac{2\pi}{R}\sum_{q=0}^{R-1}\delta(\omega-2\pi q/R)\right\} = 2\pi\delta[p]\delta(\omega) \tag{3.7}$$

where

$$F_2(m,\omega) = \sum_{n=-\infty}^{\infty} h(m-n)\, f(n) \exp[-j\omega n]$$

is the short-time Fourier transform of the synthesis filter, f(n), and, $\delta(\omega)$ is the unit impulse. Eqn. (3.7) requires that $F_2(m,\omega)$ have the property that

$$F_2(m,\omega) = \begin{cases} R & \text{for } m = 0 \text{ and } \omega = 0 \\ 0 & \text{for } m = \pm M, \pm 2M, \ldots \\ 0 & \text{for } \omega = 2\pi/R,\ 4\pi/R, \ldots,\ 2\pi(R-1)/R\ . \end{cases} \tag{3.8}$$

This formulation of the condition, which must be satisfied by the analysis and synthesis filters, will be particularly useful for dealing with the problem of linear filtering based on discrete short-time Fourier analysis considered in Section 3.4.

## 3.3 SHORT-TIME FOURIER ANALYSIS AND SYNTHESIS BASED ON THE FFT

One difficulty in implementing systems based on short-time Fourier analysis has been the rapid increase in the amount of computation required for the analysis and synthesis as the number of frequency samples becomes large. This section develops computationally efficient implementations for discrete short-time Fourier analysis and synthesis using the FFT algorithm [Portnoff].

In order to simplify notation in this section, the discrete short-time Fourier transform (for R = 1) is denoted as

$$X_k[n] = X_2(n, 2\pi k/M) \tag{3.9}$$

and the complex exponentials corresponding to the Mth roots of unity are denoted as

$$W_M^k = \exp[j2\pi k/M] \ . \tag{3.10}$$

Thus, the definition of the discrete short-time Fourier transform of the sequence x(n), for R = 1, becomes

$$X_k[n] = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m)\, W_M^{-mk} \ . \tag{3.11}$$

### 3.3.1 IMPLEMENTATION OF THE SHORT-TIME FOURIER ANALYZER

If the number of frequency samples M is chosen to be a highly composite number (usually an integral power of 2), then the FFT algorithm can be employed to compute efficiently the short-time Fourier transform $X_k[n]$ defined by eqn. (3.11). Because eqn. (3.11) does not have the form of a discrete Fourier transform (DFT), it cannot be directly computed with the FFT algorithm. The limits on the summation are given as infinite, but in practice are finite and determined by the length of h(n). By recognizing $X_k[n]$, for fixed n, as samples, equally spaced in frequency, of the (continuously-valued) partial Fourier transform of $x(m)h(n-m)$, $X_k[n]$ can be obtained by time-domain aliasing $x(m)h(n-m)$ and then computing the DFT of the aliased sequence.

Substituting $s = m - n$ into eqn. (3.11) gives

$$X_k[n] = \sum_{s=-\infty}^{\infty} x(n+s)h(-s)W_M^{-(n+s)k}$$

$$= W^{-nk} \sum_{s=-\infty}^{\infty} x(n+s)h(-s)W^{-sk}$$

which can be rewritten as

$$X_k[n] = W_M^{-nk} \sum_{l=-\infty}^{\infty} \sum_{m=0}^{M-1} x(n+lM+m)h(-lM-m)W_M^{-(lM+m)k}$$

by taking $s = lM + m$ for $m = 0, 1, \ldots, M-1$ and $l = \infty, \ldots, -1, 0, +1, \ldots, \infty$. Interchanging the orders of summation and using $W_M^M = 1$ gives

$$X_k[n] = W_M^{-nk} \sum_{m=0}^{M-1} \sum_{l=-\infty}^{\infty} x(n+lM+m)\, h(-lM-m)\, W_M^{-mk}$$

or

$$X_k[n] = W_M^{-nk} \sum_{m=0}^{M-1} \underset{\sim}{x}_m[n]\, W_M^{-mk} \tag{3.12}$$

where

$$\underset{\sim}{x}_m[n] = \sum_{l=-\infty}^{\infty} x(n+lM+m)\, h(-lM-m) \; . \tag{3.13}$$

The expression

$$\underset{\sim}{X}_k[n] = \sum_{m=0}^{M-1} \underset{\sim}{x}_m[n]\, W_M^{-mk}$$

is recognized as the partial DFT of the M-point (in m) sequence $\underset{\sim}{x}_m[n]$ for fixed n and can, therefore, be computed directly with the FFT algorithm once $\underset{\sim}{x}_m[n]$ has been formed.

In addition to the computational savings gained by computing the short-time Fourier transform using the FFT, further savings may be gained by avoiding the complex multiplications by $W_M^{-nk}$ in eqn. (3.12). Observing that $X_k[n]$ is given by

$$X_k[n] = W_M^{-nk}\, \underset{\sim}{X}_k[n]$$

where $\underset{\sim}{X}_k[n]$ is the partial DFT, with respect to m, of $\underset{\sim}{x}_m[n]$, the property that a circular shift in one domain corresponds to multiplication by a complex exponential in the other domain can be exploited. Thus, by circularly shifting $\underset{\sim}{x}_m[n]$, in m, prior to computing its DFT, the multiplications by $W_M^{-nk}$ are avoided. Specifically, eqn. (3.12) can be

rewritten as

$$X_k[n] = \sum_{m=0}^{M-1} \underset{\sim}{x}_{((m-n))_M}[n] W_M^{-mk}$$

or

$$X_k[n] = \sum_{m=0}^{M-1} x_m[n] W_M^{-mk} \tag{3.14}$$

where

$$x_m[n] = \underset{\sim}{x}_{((m-n))_M}[n] \tag{3.15}$$

and $((n))_M$ denotes "n reduced modulo M."

Based on the preceding analysis, the procedure for computing the discrete short-time Fourier transform coefficients $X_k[n]$ at a particular value of n is the following. Referring to Figure 3.3, the input data sequence considered as a function of the dummy index m is multiplied by the window $h(n-m)$ (in practice, $h(n)$ is often zero phase, in which case $h(n-m) = h(m-n)$). $h(n)$ is assumed to be of finite duration and, in fact, chosen to have length equal to an even multiple of M, plus one. The resulting weighted sequence is partitioned into sections each of length M such that $x(m)\big|_{m=n}$ is the zeroth sample of one of the sections. The resulting M-point subsequences, denoted by $x_m^{(l)}[n]$ for $0 \leq m \leq M-1$, are then added together to form

$$\underset{\sim}{x}_m[n] = \sum_l x_m^{(l)}[n], \quad \text{for } m = 0, 1, \ldots, M-1.$$

$\underset{\sim}{x}_m[n]$ is circularly shifted (in m) by n samples according to eqn. (3.15) to obtain $x_m[n]$ and, its partial DFT is computed by means of the FFT algorithm

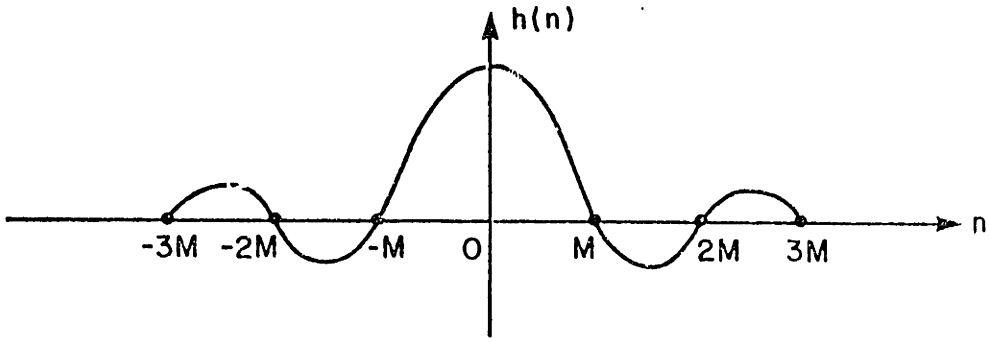



Figure 3.3(a)
Typical Unit-Sample Response for the Analysis Filter

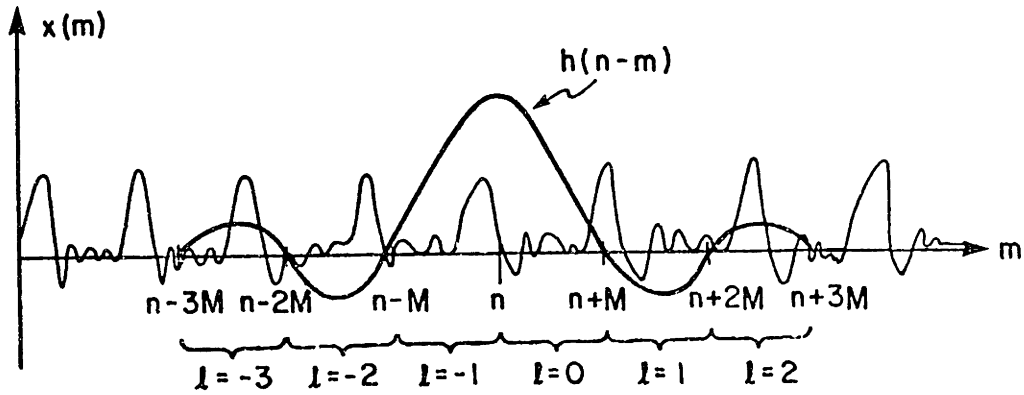



Figure 3.3 (b)
Unit-Sample Response h(n) Shifted and Superimposed on Data x(m)

to give the desired $X_k[n]$ according to eqn. (3.14).

### 3.3.2 IMPLEMENTATION OF THE SHORT-TIME FOURIER SYNTHESIZER

Let $Y_k[sR]$ denote the samples of data available to the synthesizer and let $y(n)$ denote the time sequence to be synthesized. The discrete short-time Fourier synthesis formula

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR)\, Y_k[sR]\, W_M^{nk} \qquad (3.16)$$

can be interpreted according to the discussion in Section 3.2 as interpolating each of the M sequences $Y_k[sR]$ for $k = 1,\ldots,M-1$, by 1:R, modulating each by $W_M^{nk}$ and summing the resulting signals. Clearly, this procedure could be implemented directly [Schafer and Rabiner, 1973a], however, it is computationally intractable for large values of M.

A synthesis procedure will now be formulated, which, for a highly composite number M, permits $y(n)$ to be computed from the samples $Y_k[sR]$ using the FFT algorithm. In addition to the computational savings afforded by employing the FFT, the number of computations required to perform the 1:R interpolation is reduced by the factor M.

Assume that $f(n)$ is the unit-sample response of a 1:R FIR interpolating filter with length $2QR + 1$. The synthesis formula (3.16) then becomes

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=L^-}^{L^+} f(n-sR)\, Y_k[sR]\, W_M^{nk} \qquad (3.17)$$

where the limits on the inner sum, determined by the length of $f(n)$, are

$$L^+(n) = \lceil n/R \rceil + Q$$

$$L^-(n) = \lceil n/R \rceil - Q + 1$$

and where $\lceil N \rceil$ means "the largest integer contained in N." Since the limits on both sums are finite, the order of summation can be interchanged to give

$$y(n) = \sum_{s=L^-}^{L^+} f(n-sR) \left( \frac{1}{M} \sum_{k=0}^{M-1} Y_k[sR]\, W_M^{nk} \right) \qquad (3.18)$$

or

$$y(n) = \sum_{s=L^-}^{L^+} f(n-sR)\, y_n[sR] \qquad (3.19)$$

where

$$y_n[sR] = \frac{1}{M} \sum_{k=0}^{M-1} Y_k[sR]\, W_M^{nk} \,. \qquad (3.20)$$

Thus, for fixed values of s, $y_n[sR]$ is the inverse partial DFT of $Y_k[sR]$ with respect to k, and can, therefore, be computed by the FFT algorithm. It is important to observe that $y_n[sR]$ is periodic in n with period M. Since the FFT only computes values of $y_n[sR]$ for one period ($n = 0, 1, \ldots, M-1$), the subscript n in eqn. (3.20) is interpreted as reduced modulo M.

The synthesis procedure implied by eqns. (3.19) and (3.20) can be interpreted as follows. Let $Y_k[n]$ represent the interpolated values of $Y_k[sR]$ obtained by 1:R interpolation of $Y_k[sR]$ with $f(n)$, that is,

$$Y_k[n] = \sum_{s=L^-}^{L^+} f(n-sR)\, Y_k[sR] \quad . \tag{3.21}$$

Consider the two-dimensional "net" shown in Figure 3.4. The points on the net represent the discrete set of points on which $Y_k[n]$ is defined. The horizontal direction represents time and the vertical frequency. The points corresponding to the values $Y_k[sR]$, which are available to the synthesizer, i.e., every Rth column are indicated by shading. Let $y_m[n]$ denote the inverse partial DFT of $Y_k[n]$, with respect to k:

$$y_m[n] = \frac{1}{M} \sum_{k=0}^{M-1} Y_k[n] W_M^{mk} \quad . \tag{3.22}$$

$y_m[n]$ is defined on the net shown in Figure 3.5 and the samples $y_m[sR]$ are defined on the shaded points. Inverse transforming eqn. (3.21) with respect to k shows that $y_m[n]$ corresponds to the interpolated values of $y_m[sR]$ obtained by 1:R interpolation of $y_m[sR]$ with f(n), that is,

$$y_m[n] = \sum_{s=L^-}^{L^+} f(n-sR)\, y_m[sR] \quad . \tag{3.23}$$

A comparison of eqns. (3.19) and (3.23) indicates that the values of y(n) are the values of $y_m[n]$ for $m \equiv n \bmod M$, which correspond to the points in Figure 3.5 on the "helical" path, $m \equiv n \bmod M$. The operation defined by eqn. (3.19) is, therefore, interpreted as interpolating $y_m[sR]$ to obtain the unknown values of $y_m[n]$, but only those values of $y_m[n]$ on the path $m \equiv n \bmod M$ that are the values of y(n).

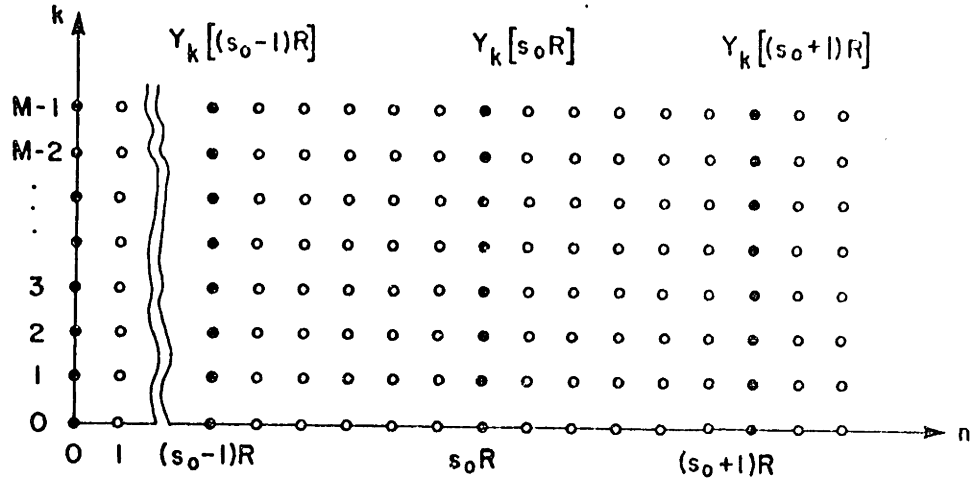



Figure 3.4

Net on which $Y_k[n]$ is Defined

Shaded points represent the values of $Y_k[sR]$

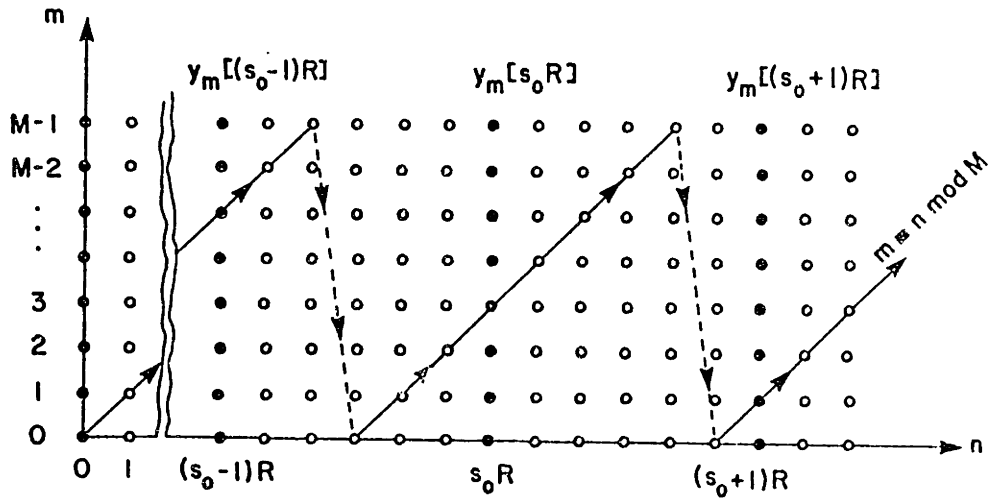



Figure 3.5

Net on which $y_m[n]$ is Defined

Shaded points represent the values of $y_m[sR]$

Values along the path $m \equiv n \bmod M$ are $y(n) = y_n[n]$

The implementation of the synthesis procedure is, therefore, as follows. First, the values of $y_m[sR]$ are obtained by inverse transforming $Y_k[sR]$ with respect to k using the FFT. The values of y(n) are then obtained by interpolating $y_m[sR]$ according to eqn. (3.19). Notice that for each value of y(n), 2Q values of $y_m[sR]$ are required. In fact, for R consecutive values of y(n), these values are obtained from the same 2Q columns. Thus, it is natural to compute y(n) in records of length R. For each output value, imagine a mask that extracts 2Q values of $y_m[sR]$, as shown in Figure 3.6. These values are then processed according to eqn. (3.19) to compute y(n). Successive output values are obtained by shifting the mask one sample at a time along the path $m=n \bmod M$ and repeating the process.

## 3.4 LINEAR FILTERING BASED ON DISCRETE SHORT-TIME FOURIER ANALYSIS

Section 3.2 showed that a sequence could be represented with no redundancy by the discrete-short-time Fourier transform. This section now considers the problem of linearly filtering a sequence by multiplying its discrete short-time Fourier transform by samples of the time-varying frequency response of the filter. If the sampled-transform implementation is to be equivalent to the formulation of Section 2.6, then the sampling rates in both time and frequency must be sufficiently high to represent both the short-time Fourier transform and the time-varying frequency response with no aliasing in either time or frequency.

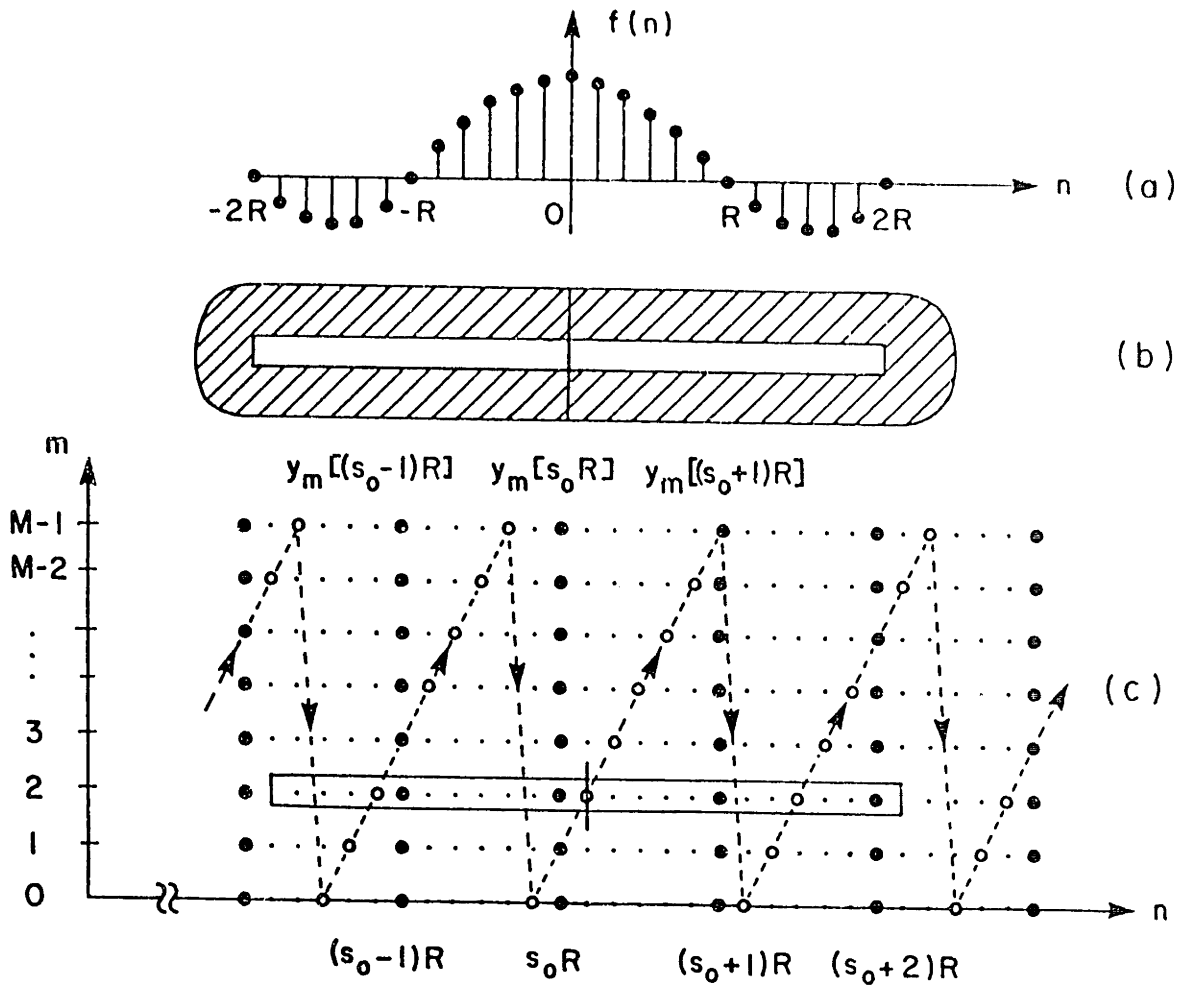



Figure 3.6

(a) Typical Unit-Sample Response for a 1:R FIR Digital Interpolating Filter f(n)

(b) Mask to Extract Values of $y_m[sR]$ to be Interpolated Using f(n)

(c) Net Associated with $y_m[n]$

- • indicates points representing $y_m[sR]$
- o Indicates points representing $y(n) = y_n[n]$

### 3.4.1 LINEAR TIME-VARYING FILTERS

Let $X(sR, k\Omega_0)$ represent samples of the short-time Fourier transform of $x(n)$ as defined by eqn. (3.1) and let $T_2(sR, k\Omega_0)$ represent samples of the time-varying frequency response of the linear time-varying system with unit-sample response $t(n, m)$. Furthermore, assume that

$$T_1(\psi, m) = \sum_{n=-\infty}^{\infty} t(n, m) \exp[-j\psi n] \tag{3.24}$$

is bandlimited in $\psi$ and time limited in $m$ so that $t(n, m)$ can be represented, without aliasing, by samples of $T_2(n, \omega)$.

Define $Y_2(sR, k\Omega_0)$ as the product

$$Y_2(sR, k\Omega_0) = T_2(sR, k\Omega_0)\ X_2(sR, k\Omega_0) \tag{3.25}$$

and $y(n)$ as

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR)\, Y(sR, k\Omega_0) \exp[j\Omega_0 kn] \ . \tag{3.26}$$

Thus,

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR)\, T_2(sR, k\Omega_0)\, X_2(sR, k\Omega_0) \exp[j\Omega_0 kn] \tag{3.27}$$

and $y(n)$ corresponds to the output of the system depicted in Figure 3.7.

The overall unit-sample response of this system, denoted $\tilde{t}(n, m)$, is given by

$$\tilde{t}(n, m) = \sum_{p=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR-n+m)\, t(sR, m-pM) \tag{3.28}$$

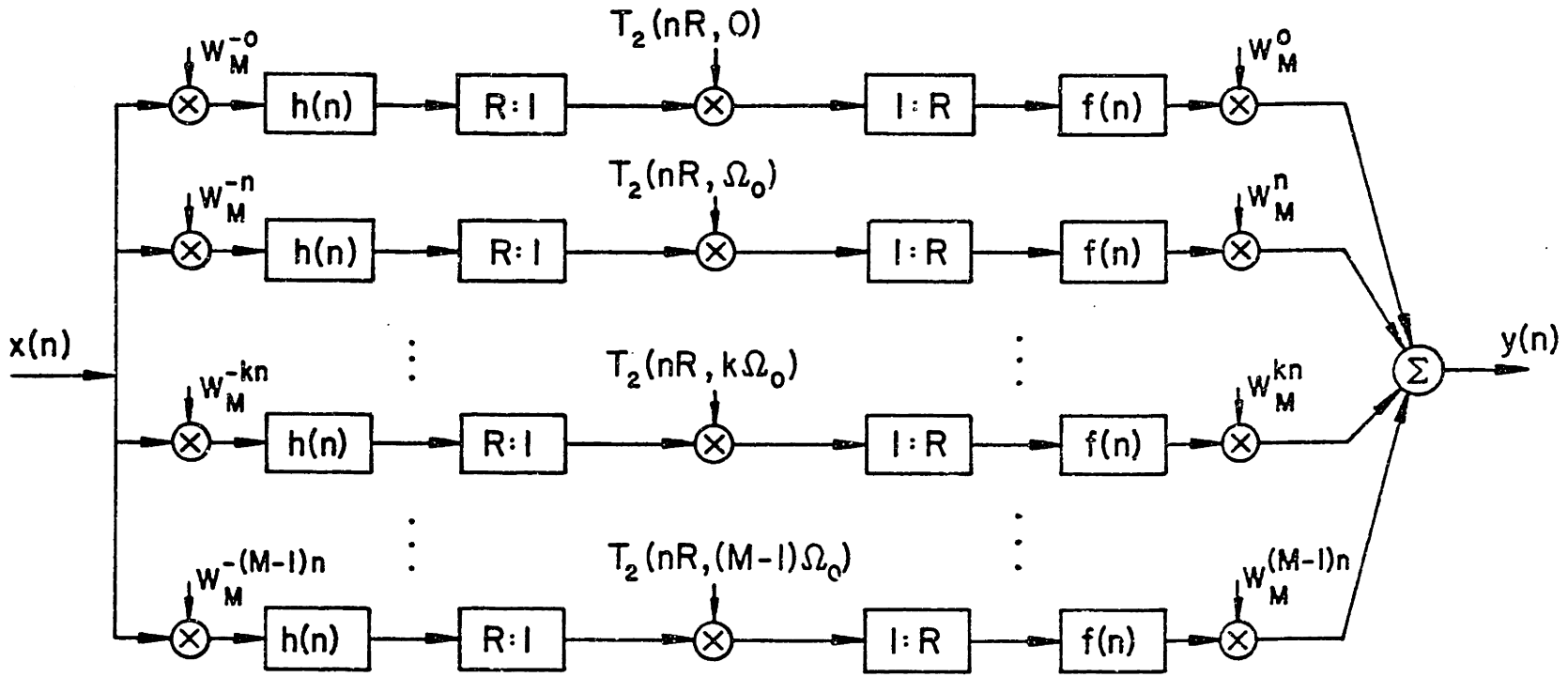



Figure 3.7

Filterbank Analogue for Linear Time-Varying Filtering
Based on Short-Time Fourier Analysis

and the output of the system, $y(n)$, is given by

$$y(n) = \sum_{m=-\infty}^{\infty} \tilde{t}(n,m)\, x(n-m) \tag{3.29}$$

Eqns. (3.28) and (3.29) follow from substituting the definition (3.1) of $X(sR, k\Omega_0)$ into eqn. (3.26) to get

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR)\, T_2(sR, k\Omega_0)$$

$$\times \sum_{m=-\infty}^{\infty} h(sR-m)\, x(m) \exp[-j\Omega_0 km] \exp[j\Omega_0 kn]$$

$$= \sum_{s=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} f(n-sR)\, h(sR-m)$$

$$\times \frac{1}{M} \sum_{k=0}^{M-1} T_2(sR, k\Omega_0) \exp[j\Omega_0 k(n-m)]\, x(m) \ .$$

Now letting $m' = n-m$ and interchanging the order of summation gives

$$y(n) = \sum_{m'=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR-m'-n)\, t(sR, m'-pM)\, x(n-m')$$

$$= \sum_{m'=-\infty}^{\infty} \tilde{t}(n,m')\, x(n-m') \ .$$

The conditions on the filters and sampling rates such that $\tilde{t}(n,m)$ defined by eqn. (3.28) for the sampled transform implementation is identical to $\bar{t}(n,m)$ defined by eqn. (2.53) for the non-sampled implementation become apparent by expressing the partial Fourier transform $T_1(\psi, m)$ in terms of $\tilde{T}_1(\psi, m)$. Transforming eqn. (3.28) with respect to n

gives

$$\hat{T}_1(\psi, m) = \sum_{n=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR-n+m)\, t(sR, m-pM) \exp[-j\psi n]$$

and, letting $r = n-sR$,

$$\hat{T}_1(\psi, m) = \sum_{r=-\infty}^{\infty} \sum_{p=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} f(r)\, h(m-r)\, t(sR, m-pM) \exp[-j(r+sR)]$$

$$= \sum_{r=-\infty}^{\infty} f(r)\, h(m-r) \exp[-j\psi r] \times \sum_{p=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} t(sR, m-pM) \exp[-j\psi sR]$$

hence

$$\hat{T}_1(\psi, m) = F_2(m, \psi)\, \frac{1}{R} \sum_{q=0}^{R-1} \sum_{p=-\infty}^{\infty} T_1(\psi - 2\pi q/R, m-pM) \tag{3.30}$$

where $T_1(\psi, m)$ is the partial Fourier transform of $t(n, m)$ and $F_2(m, \psi)$ is the short-time Fourier transform of $f(n)$ given by

$$F_2(m, \psi) = \sum_{r=-\infty}^{\infty} h(m-r)\, f(r) \exp[-j\psi r]$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\psi+\varphi)\, H(\varphi) \exp[j\varphi m]\, d\varphi \,. \tag{3.31}$$

Eqn. (3.30) states that $T_1(\psi, m)$ is aliased in both time and frequency, then windowed by $F_2(m, \psi)$ to produce $\hat{T}_1(\psi, m)$.

For the sampled-transform implementation to be equivalent to the non-sampled implementation of Section 2.6, $\hat{T}_1(\psi, m)$, given by eqn. (3.30), must be identical to $\hat{T}_1(\psi, m)$, given by eqn. (2.53). To prevent distortion due to aliasing, the number of frequency samples, M, must be at least as

great as the duration of $T_1(\psi, m)$ in $m$, and the temporal sampling frequency, $2\pi/R$, must be greater than the bandwidth of $T_1(\psi, m)$ in $\psi$. Further, to eliminate the images of $T_1(\psi, m)$ in eqn. (3.30), M must be, in general, at least as great as the duration of $F_2(m, \psi)$ in m and, $2\pi/R$ must be greater than the bandwidth of $F_2(m, \psi)$ in $\psi$. The latter condition is not necessary for the special case of $T_1(\psi, m)$ and $F_2(\psi, m)$ having structures such that regions of zero of $F_2(m, \psi)$ exactly cancel the images of $T_1(\psi, m)$ in eqn. (3.30). One example of this situation is the implementation of fast convolution for linear time-invariant filters, treated in the following section.

### 3.4.2 FAST CONVOLUTION

To conclude this chapter, the methods of fast convolution [Stockham] are considered as special cases of linear filtering based on discrete short-time Fourier analysis. If a signal is processed by multiplying its short-time Fourier transform by a time-invariant frequency response, and a new signal synthesized from the product, then if the analysis and synthesis filters have appropriate rectangular unit-sample responses, the overall processing is equivalent (but not identical) to the overlap-save or overlap-add method of fast convolution, depending upon the lengths of the unit-sample responses of the analysis and synthesis filters.

In particular, if the length of the filter unit-sample response to be implemented is L so that

$$t(n,m) = \begin{cases} t(m) & \text{for } 0 \leq m < L \\ 0 & \text{otherwise} \end{cases} \tag{3.32}$$

then the overlap-save technique results when

$$h(m) = \begin{cases} 1 & \text{for } 0 \leq m < M \\ 0 & \text{otherwise} \end{cases}$$

$$f(m) = \begin{cases} 1 & \text{for } -R < m \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

and the processing is implemented as discussed in Section 3.3, with the parameters L, M and R satisfying the constraint

$$M \geq L + R - 1 . \tag{3.34}$$

In the terminology associated with the fast convolution technique, L is the length of the filter unit-sample response, M is the length of each input-data section which is equal to the DFT size, and R is the number of "good" points per output section which is equal to the spacing between overlapping input-data sections. Similarly, the overlap-add technique results when the designs for h(m) and f(m) are interchanged so that

$$h(m) = \begin{cases} 1 & \text{for } -R < m \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f(m) = \begin{cases} 1 & \text{for } 0 \leq m < M \\ 0 & \text{otherwise} \end{cases} \tag{3.35}$$

and the processing is again implemented according to the Section 3.3 with the constraint (3.34). Here R is the length of each of the contiguous input-data sections which is equal to the spacing between overlapping output sections and M is the length of each of the output-data sections which is equal to the DFT size.

The difference between the formulation based on short-time Fourier analysis and the conventional formulation is that in the conventional formulation the analogous filterbank channel signals are bandpass signals as illustrated in Figure 3.8a, whereas, in the formulation based on short-time Fourier analysis, the channel signals are low-pass signals as illustrated in Figure 3.8b. Because the modulation operations on the channel signals are actually implemented by circular shifts as discussed in Section 3.3.1, the difference between the two formulations amounts to a difference in indexing schemes.

Linear convolution with $t(m)$ is now shown to result if the analysis and synthesis filter pairs (3.33) or (3.35) are used with parameters satisfying (3.34). For a linear time-invariant system with $t(n,m) = t(m)$, $T_1(\psi,m)$ and $\hat{T}_1(\psi,m)$ become

$$T_1(\psi,m) = 2\pi\delta(\psi)\,t(m)$$

and

$$\hat{T}_1(\psi,m) = F_2(m,\psi)\cdot\frac{2\pi}{R}\left\{\sum_{q=0}^{R-1}\delta(\psi-2\pi q/R)\right\}\left\{\sum_{p=-\infty}^{\infty}t(m-pM)\right\} \qquad (3.36)$$

For the overlap-save technique, the filter designs (3.33) result in the short-time Fourier transform of $f(m)$ given by

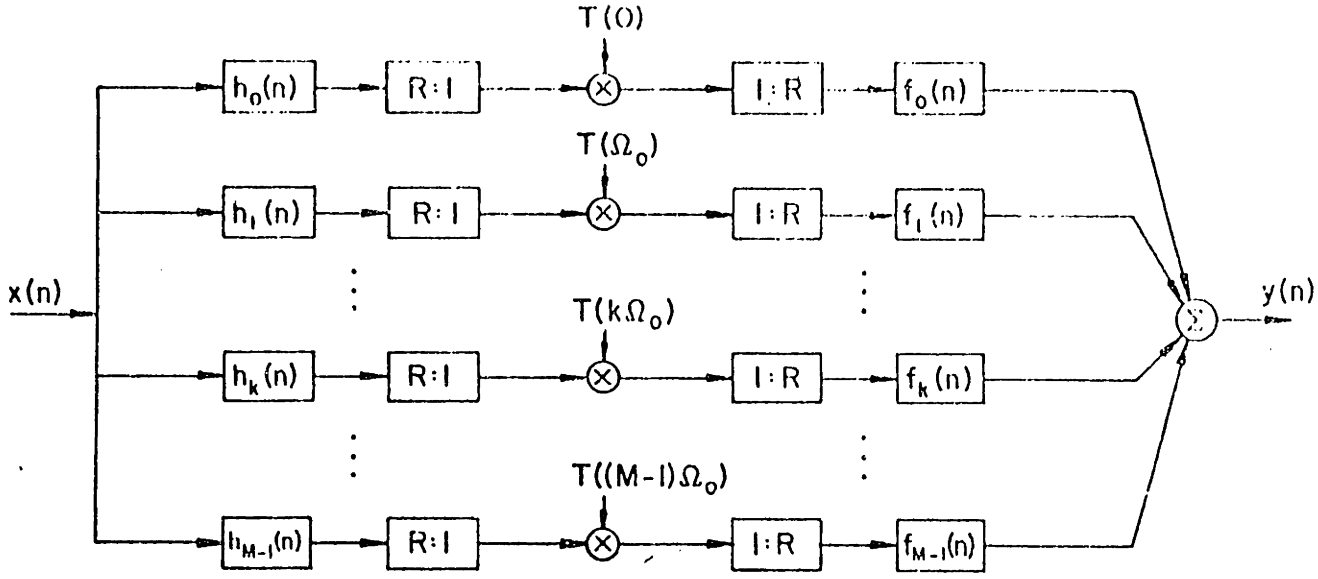



Figure 3.8(a)

Filterbank Analogue for Conventional Method of Fast Convolution

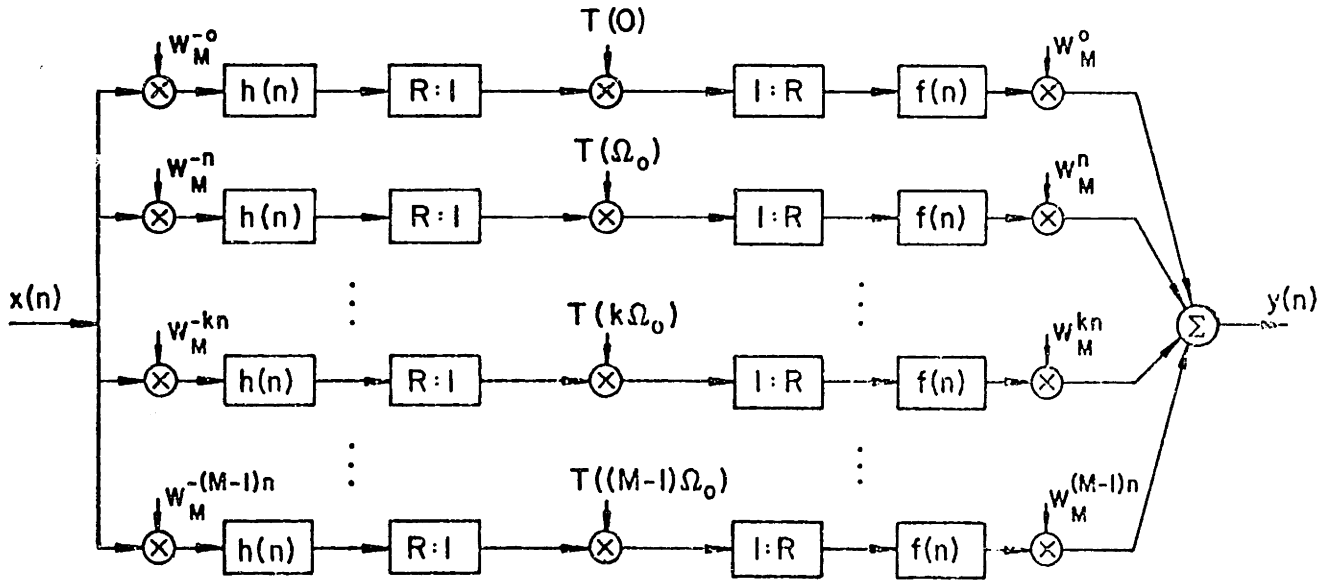



Figure 3.8(b)

Filterbank Analogue for Short-Time Fourier Transform Method of Fast Convolution

$$F_2(m,\psi) = \begin{cases} e^{j\psi(R-1-m)/2} \cdot \dfrac{\sin[\psi(m+R)/2]}{\sin(\psi/2)} & \text{for } -R < m \leq 0 \\ e^{j\psi(R-1)/2} \cdot \dfrac{\sin[\psi R/2]}{\sin(\psi/2)} & \text{for } 0 \leq m \leq M-R \\ e^{j\psi(M-1-m)/2} \cdot \dfrac{\sin[\psi(M-m)/2]}{\sin(\psi/2)} & \text{for } M-R \leq m < M \\ 0 & \text{otherwise} \end{cases} \tag{3.37}$$

By considering $\tilde{T}_1(\psi,m)$ for each of the four conditions in eqn. (3.37), $\tilde{T}_1(\psi,m)$ can be seen to reduce to $T_1(\psi,m)$. For $m$ in the range $-R < m \leq 0$ or $M-R \leq m < M$, $\tilde{T}_1(\psi,m) = 0$ because $t(m) = 0$ for $m < 0$ or $L \leq m$ and because $M$, $L$ and $R$ satisfy (3.34). For $m$ in the range $0 \leq m \leq M-R$, $\tilde{T}_1(\psi,m) = 2\pi\delta(\psi)t(m)$ because $F_2(m, 2\pi q/R) = R$ for $q = 0$ and $F_2(m, 2\pi q/R) = 0$ for $0 < q < R$. For all other choices of $m$ $F_2(m,\psi) = 0$, hence $\tilde{T}_1(\psi,m) = T_1(\psi,m)$ for all $\psi$ and $m$. The overall unit-sample response of the system $\tilde{t}(n,m)$ is exactly $t(m)$ and the desired linear convolution is achieved.

For the overlap-add technique, the designs for $h(m)$ and $f(m)$ are interchanged, thus, the short-time Fourier transform of $f(m)$ for the overlap-add technique is just the Fourier transform of $f(m)$ for the overlap-save technique (3.37) with $\psi$ replaced by $-\psi$ and multiplied by $\exp[-j\psi n]$. By the previous argument $\tilde{t}(n,m) = t(m)$ and the desired linear convolution is again achieved.

# CHAPTER 4

# QUASI-STATIONARY REPRESENTATION OF SAMPLED SPEECH SIGNALS

## 4.1 INTRODUCTION

In this chapter, a mathematical representation for the sampled speech signal is formulated. This representation provides the framework necessary to describe and implement time-scale compression and expansion of speech. Not only must such a representation characterize the speech waveform, but it must also have the property that simple time-scaling of the parameters of the representation corresponds to changing the rate of the speech. Therefore, the development will begin with an appropriate model of the speech signal in order to provide a mathematical definition of rate-changed speech.

## 4.2 A QUASI-STATIONARY REPRESENTATION FOR SPEECH SIGNALS

The generally accepted engineering model for the production of speech signals is illustrated in Figure 4.1. According to this model, samples of the speech waveform, denoted by $x(n)$, are assumed to be the output of a linear time-varying filter that approximates the transmission characteristics of the vocal tract and the spectral characteristics of the glottal pulse. For voiced speech, the filter is driven by a quasi-periodic train of unit samples, $v(n)$, such that the spacing between the unit samples corresponds to the pitch, or fundamental, period of the speech. For unvoiced speech, the filter is driven by a stationary random sequence, $u(n)$ with a flat power spectrum, i.e., white noise.

The input-output behavior of the linear time-varying filter depicted in Figure 4.1 is completely characterized by its time-varying unit-sample response $t(n,m)$, or equivalently, by its time-varying frequency response $T_2(n,\omega)$. Recall, from Chapter 2, that $t(n,m)$ represents the response of the system at time sample $n$ to a unit sample applied $m$ samples earlier, whereas, $T_2(n,\omega)$ represents the response of the system at time sample $n$ to the complex exponential $\exp[j\omega n]$. Furthermore, the time variation, or nonstationarity, of the system is manifested by the dependence of the functions $t(n,m)$ and $T_2(n,\omega)$ on the index $n$.

In the speech production model, the nonstationarity of $t(n,m)$ corresponds to the movement of the physical articulators and is usually relatively slow compared to the time variation of the input and output waveforms that correspond to acoustic signals. Because the vocal tract is,

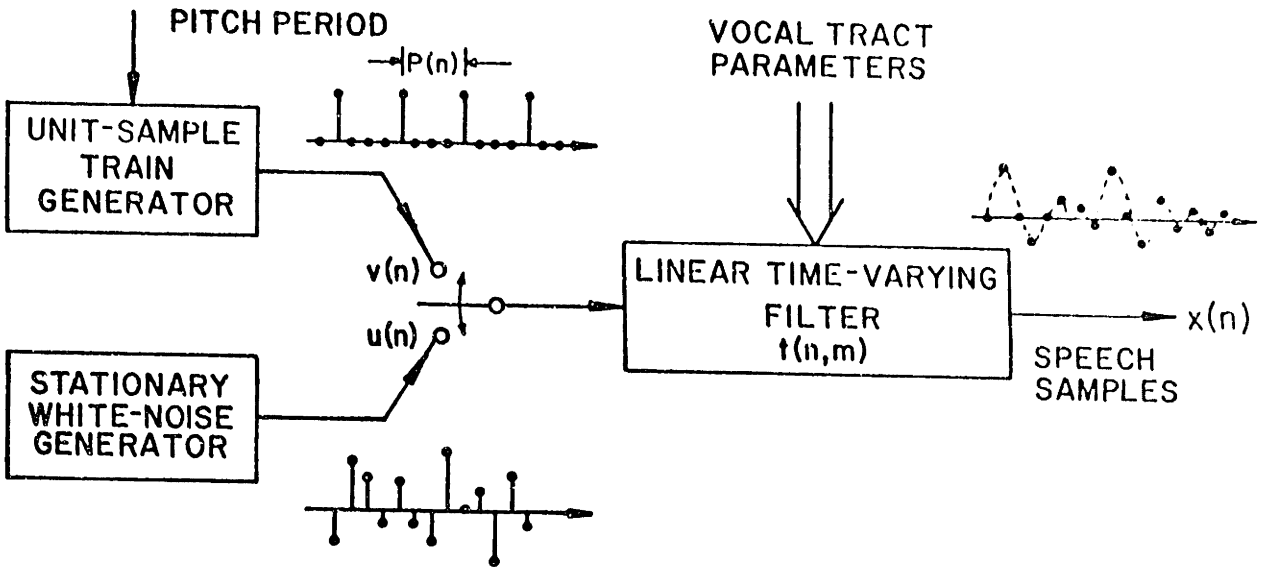



Figure 4.1
Terminal-Analogue Model of the Vocal System
(after Schafer and Rabiner, 1975)

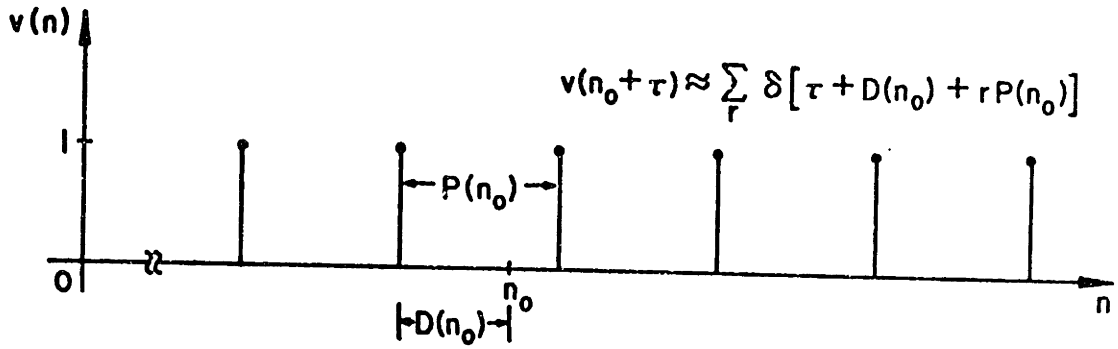



Figure 4.2
Quasi-Periodic Unit-Sample Train

in essence, a lossy acoustic resonator, its impulse response can be modelled as finite duration and causal. Moreover, the time variation of the vocal tract is sufficiently slow compared to the decay rate of its impulse response that the system can be modelled as almost stationary for the duration of its memory. Such a system, which is approximately stationary for the duration of its memory, will be termed a "quasi-stationary system."

### 4.2.1 HARMONIC REPRESENTATION OF VOICED SPEECH

According to Figure 4.1, voiced speech is modelled as the output of $t(n,m)$ driven by the quasi-periodic unit-sample train $v(n)$. The term "quasi periodic" is used to indicate that $v(n)$ has a locally periodic behavior. Specifically, if $v(n)$, illustrated in Figure 4.2, is expressed relative to a sliding time frame as $v(n_0+\tau)$, representing, for a fixed value of $n_0$ and small $|\tau|$, the local behavior of $v(n)$, then $v(n_0+\tau)$ is periodic in $\tau$ for $(n_0+\tau)$ in the neighborhood about $n_0$ for which $v(n)$ is modelled as periodic.

A harmonic representation for voiced-speech signals will now be formulated by representing the excitation, $v(n)$, as a sum of harmonically related complex exponentials and using this representation in the superposition sum with $t(n,m)$. Referring to Figure 4.2, we let $P(n_0)$ denote the local pitch period in the neighborhood of $n_0$ and let $D(n_0)$ denote the number of samples to the sample $n_0$ from the unit sample arriving most recently before, or at the sample $n_0$. $v(n)$ can, therefore, be

represented locally as

$$v(n_0+\tau) \approx \sum_{r=-\infty}^{\infty} \delta[\tau+D(n_0)+rP(n_0)] \tag{4.1}$$

for small $|\tau|$. Equivalently, $v(n_0+\tau)$ can be represented as the sum of harmonically related complex exponentials

$$v(n_0+\tau) \approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp[j2\pi k(D(n_0)+\tau)/P(n_0)] \tag{4.2}$$

or,

$$v(n_0+\tau) \approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} \exp[jk(\phi(n_0)+\Omega(n_0)\tau+\phi_0)] \tag{4.3}$$

where

$$\Omega(n) = 2\pi/P(n) \tag{4.4a}$$

$$\phi_0 = \Omega(0)D(0) \tag{4.4b}$$

$$\phi(n) = \Omega(n)D(n) + 2\pi I(n) - \phi_0 \tag{4.4c}$$

and $I(n)$ is an integer whose value depends on n. That eqns. (4.1) and (4.2) are equivalent, follows directly from evaluating the sum in eqn. (4.2), for fixed $n_0$, as the sum of a finite number of terms in a geometric series.

The quantity $\Omega(n)$, referred to as the "instantaneous frequency" of the fundamental, can assume values in the range

$$0 \leq \Omega(n) < 2\pi \tag{4.5}$$

because the pitch period, $P(n)$, is defined as a positive integer and, for speech, is greater than unity. Furthermore, because $v(n)$ is quasi periodic, both $P(n)$ and $\Omega(n)$ are slowly varying, so that

$$P(n_0 + \tau) \approx P(n_0) \qquad \text{for small } |\tau| \tag{4.6a}$$

and

$$\Omega(n_0 + \tau) \approx \Omega(n_0) \qquad \text{for small } |\tau| \, . \tag{4.6b}$$

The constant phase angle, $\phi_0$, is introduced as a convenience to preserve the time origin under rate-change modifications. Note that $\phi_0$ is zero if one of the unit samples of $v(n)$ arrives at $n = 0$.

The quantity $\phi(n)$ is referred to as the "instantaneous phase" of the fundamental. The unspecified additive multiple of $2\pi$, in eqn. (4.4c) can be specified by requiring that the value of the exponent in eqn. (4.3) be uniquely defined, for each value of $n = n_0 + \tau$. Thus, for $n = n_0 + \tau$, we require that

$$\phi(n) + \phi_0 \approx \phi(n_0) + \Omega(n_0)\tau + \phi_0 \tag{4.7}$$

or

$$\phi(n_0 + \tau) \approx \phi(n_0) + \Omega(n_0)\tau \tag{4.8}$$

for small $|\tau|$. Further, setting $\tau = -1$ in eqn. (4.8) leads to the property that the instantaneous frequency is the first (backwards) difference of the instantaneous phase, i.e.,

$$\Omega(n) \approx \phi(n) - \phi(n-1) \, . \tag{4.9}$$

Furthermore, since

$$\phi(n) \approx \phi(n-1) + \Omega(n) \tag{4.10}$$

$\phi(n)$ will now be <u>defined</u> explicitly as

$$\phi(n) = \begin{cases} \sum_{r=1}^{n} \Omega(r) & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=0}^{n+1} -\Omega(r) & \text{for } n < 0 \end{cases} \qquad (4.11)$$

and eqn. (4.11) will be taken as the definition of $\phi(n)$. Note that, from (4.9) and (4.5), $\phi(n)$ has the property

$$0 \leq \phi(n) - \phi(n-1) < 2\pi$$

and is, therefore, an "unwrapped phase" angle rather than a principal value. If $v(n)$ is periodic with period P for all time, then $\Omega(n) = 2\pi/P$ is a constant and, $\phi(n)$ becomes $2\pi n/P$.

The excitation, $v(n)$, for voiced speech will, therefore, be modelled as the sum of harmonically related complex exponentials

$$v(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp[jk(\phi(n) + \phi_0)] \qquad (4.12)$$

where $\phi(n)$ is defined by eqn. (4.11), $\Omega(n) = 2\pi/P(n)$, and $\phi_0 = \Omega(0)D(0)$. For a periodic unit-sample train with period P the representation (4.12) becomes, simply,

$$v(n) = \frac{1}{P} \sum_{k=0}^{P-1} \exp[jk(\frac{2\pi n}{P} + \phi_0)] \ .$$

The harmonic representation (4.12) is, in fact, more general than the time-domain representation (4.1). In addition to representing periodic sequences, the harmonic representation also represents aperiodic sequences with periodic envelopes, such as those obtained by uniformly sampling periodic continuous-time waveforms when the period of the waveform is not an integer multiple of the sampling interval.

A voiced-speech signal $x(n)$, modelled as the output of $t(n,m)$ driven by $v(n)$, is given according to the superposition sum as

$$x(n) = \sum_{m} t(n,m)\, v(n-m) \ . \tag{4.13}$$

For voiced speech, the pitch, $\Omega(n)$, is assumed to be constant for the duration of the memory of $t(n,m)$. Therefore, $v(n-m)$ in eqn. (4.13) can be replaced by the local harmonic representation (4.3) to obtain

$$\begin{aligned} x(n) &= \sum_{m} t(n,m) \left\{ \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \exp[jk(\phi(n) - \Omega(n)m + \phi_0)] \right\} \\ &= \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} T_2(n, k\Omega(n)) \exp[jk(\phi(n) + \phi_0)] \ . \end{aligned} \tag{4.14}$$

Thus, the voiced-speech signal, $x(n)$, is represented as the linear combination of harmonically related complex exponentials

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) \exp[jk\phi(n)] \tag{4.15a}$$

where

$$c_k(n) = \frac{1}{P(n)} T_2(n, k\Omega(n)) \exp[jk\phi_0] \tag{4.15b}$$

and

$$\phi(n) = \begin{cases} \sum_{r=1}^{n} \Omega(r) & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=0}^{n+1} -\Omega(r) & \text{for } n < 0 \end{cases} \qquad (4.15c)$$

The quantities $c_k(n)$, referred to as the "complex harmonic amplitudes" of the speech, are slowly-varying functions of n. Because the time variation of the $c_k(n)$'s corresponds to changes in the vocal-tract geometry, the $c_k(n)$'s contain significant Fourier components only up to the order of a few tens of hertz. Thus, on the spectrum of acoustic frequencies, which extends into the tens of kilohertz, the $c_k(n)$'s are narrow-band low-pass sequences. Furthermore, a property that will be exploited in the short-time Fourier analysis of speech is that the bandwidths of the $c_k(n)$'s are much less than the fundamental frequency, $\Omega(n)$, of the speech.

### 4.2.2 SECOND MOMENT REPRESENTATION OF UNVOICED SPEECH

An unvoiced-speech signal, $x(n)$, is modelled as the output of the quasi-stationary filter $t(n,m)$ driven by the real zero-mean stationary white-noise process $u(n)$ with autocorrelation function

$$R_u(\tau) = E\{u(n+\tau)u^*(n)\}$$
$$= \sigma_u^2 \delta[\tau] \quad . \tag{4.16}$$

Clearly, $x(n)$ has zero mean and is nonstationary with its autocorrelation function given by

$$R_x(n,\tau) = E\{x(n+\tau)x^*(n)\}$$
$$= E\Big\{ \sum_q t(n+\tau,q)u(n+\tau-q) \sum_m t^*(n,m)u^*(n-m) \Big\}$$
$$= \sum_q \sum_m t(n+\tau,q)t^*(n,m)\sigma_u^2 \delta[\tau-q+m]$$
$$= \sum_m \sigma_u^2 t(n+\tau,m+\tau)t^*(n,m) \quad . \tag{4.17}$$

Since $t(n,m)$ is assumed to be a quasi-stationary system, its time variation (in $n$) is negligible over the duration of its memory (correlation time in $m$). Hence, eqn. (4.17) becomes

$$R_x(n,\tau) \approx \sum_m \sigma_u^2 t(n,m+\tau)t^*(n,m)$$

or

$$R_x(n,\tau) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} \sigma_u^2 |T_2(n,\omega)|^2 \exp[j\omega\tau]d\omega \quad . \tag{4.18}$$

A nonstationary random process, such as $x(n)$, that can be modelled as the output of a quasi-stationary linear time-varying system, will be referred to as a "quasi-stationary random process." Furthermore, as a generalization of the power spectrum for a wide-sense stationary random process, define the "time-varying power spectrum," $S_x(n,\omega)$, for the quasi-stationary random process, $x(n)$, as

$$S_x(n,\omega) = \sigma_u^2 |T_2(n,\omega)|^2 . \qquad (4.19)$$

Thus, based on the approximation (4.18), $R_x(n,\tau)$ and $S_x(n,\omega)$ can be written as the Fourier transform pair

$$R_x(n,\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega)\exp[j\omega\tau]\,d\omega \qquad (4.20)$$

and

$$S_x(n,\omega) = \sum_{\tau} R_x(n,\tau)\exp[-j\omega\tau] . \qquad (4.21)$$

An unvoiced-speech signal $x(n)$ will, therefore, be represented as a quasi-stationary random process with zero mean and characterized by its second moments, namely $R_x(n,\tau)$, or equivalently, $S_x(n,\omega)$. Moreover, this representation provides a direct relationship between the time-varying statistics of $x(n)$ and the parameters of the filter $t(n,m)$.

## 4.3 REPRESENTATION OF RATE-CHANGED SPEECH SIGNALS

As discussed in Chapter 1, the objective of this research is the development of a system to transform a given speech signal into a new speech-like signal that retains all of the features of the original speech except that it appears to have been articulated at a rate different from the original speech. If $x(n)$ denotes samples of the original speech signal, the superscript notation $x^{\beta}(n)$ will be used to denote the rate-changed signal with apparent rate of articulation multiplied by the factor $\beta$. Such modifications can be defined in terms of the representation developed in

Section 4.2; however, it is necessary, first, to clarify the notion of linear time scaling of discrete-time signals.

### 4.3.1 REPRESENTATION OF LINEARLY TIME-SCALED SEQUENCES

The representation of a discrete-time signal obtained by linearly time compressing or expanding a given discrete-time signal is fundamental to the development of the theory of time compression and expansion of speech signals. The discrete-time signals that will be of interest fall into two categories: those corresponding to samples of bandlimited continuous-time waveforms and those obtained by nonlinear transformations of such signals. A sequence $x(n)$, corresponding to samples of a continuous-time waveform, will be interpreted as samples of the continuous-time waveform $x(t)$, with the sampling interval normalized to unity. Based on this interpretation, it is meaningful to define the sequence $x(\beta n)$, corresponding to samples of $x(\beta t)$ with unity sampling interval, as $x(n)$ linearly time-scaled by $\beta$. Moreover, if $\beta$ is rational, so that it can expressed as the quotient of two integers

$$\beta = D / I ,$$

and if $x(t)$ is appropriately bandlimited, then the sequence $x(\beta n)$ can be obtained directly from $x(n)$ by the decimation / interpolation formula [Schafer and Rabiner, 1973b]

$$x(\beta n) = \sum_{r=-\infty}^{\infty} f(nD-rI)\,x(r) \qquad (4.22)$$

where $f(n)$ is a 1:I interpolating filter. A sequence $y(n)$, obtained by a nonlinear transformation of $x(n)$, above, will be linearly time-scaled by $\beta$, to obtain $y(\beta n)$, by linearly time-scaling the underlying bandlimited sequence according to eqn. (4.22).

### 4.3.2 REPRESENTATION OF RATE-CHANGED VOICED SPEECH

Let $x(n)$ represent samples of a voiced-speech signal modelled according to Section 4.2.1. The rate-changed signal, $x^{\beta}(n)$, is modelled by linearly time-scaling the time-varying parameters of the filter, $t(n,m)$, and the pitch contour, $\Omega(n)$, by the factor $\beta$. Thus, $x^{\beta}(n)$ corresponds to the output of the filter

$$t^{\beta}(n,m) = t(\beta n, m) \qquad (4.23a)$$

with time-varying frequency response

$$T_2^{\beta}(n,\omega) = T_2(\beta n, \omega) \qquad (4.23b)$$

driven by the quasi-periodic unit-sample train $v^{\beta}(n)$, with time-scaled pitch $\Omega(\beta n)$. Moreover, the excitation, $v^{\beta}(n)$, can be expressed as

$$v^{\beta}(n) = \frac{1}{P(\beta n)} \sum_{k=0}^{P(\beta n)-1} \exp[jk(\phi(\beta n)/\beta + \phi_0)] \,. \qquad (4.24)$$

To show that eqn. (4.24) represents the desired "rate-changed" unit-sample train, we must show that its pitch is the time-scaled instantaneous frequency $\Omega(\beta n)$, and the phase of its fundamental at $n = 0$ is $\phi_0$. Since $\phi(0) = 0$, the latter condition is obvious. To show the former, linearly time scale eqn. (4.8) in both $n_0$ and $\tau$ to obtain

$$\phi(\beta(n_0+\tau)) \approx \phi(\beta n_0) + \Omega(\beta n_0)\beta\tau$$

or

$$\phi(\beta(n_0+\tau))/\beta \approx \phi(\beta n_0)/\beta + \Omega(\beta n_0)\tau \ . \qquad (4.25)$$

Setting $\tau = -1$ gives

$$\Omega(\beta n) \approx \phi(\beta n)/\beta - \phi(\beta(n-1))/\beta \ , \qquad (4.26)$$

which shows that the fundamental frequency of $v^{\beta}(n)$ is indeed $\Omega(\beta n)$.

From (4.24) and (4.25), $v^{\beta}(n)$ can be represented locally as

$$v^{\beta}(n_0+\tau) \approx \frac{1}{P(\beta n_0)} \sum_{k=0}^{P(\beta n_0)-1} \exp[jk(\phi(\beta n_0)/\beta + \Omega(\beta n_0)\tau + \phi_0)] \ . \qquad (4.27)$$

Using eqn. (4.27) in the superposition sum and paralleling the development of Section 4.2.1 yields the harmonic representation for rate-changed voiced speech

$$x^{\beta}(n) = \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) \exp[jk\phi(\beta n)/\beta] \quad . \tag{4.28}$$

It is important to remember that the instantaneous phase, $\phi(n)$, in the above equations is the "unwrapped phase" angle given by eqn. (4.11), not its principal value. This distinction is important because $1/\beta$ is not, in general, an integer; thus, while an integer multiple of $2\pi$ added to the argument of the exponential in eqn. (4.28) is invisible, an integer multiple of $2\pi/\beta$ is not.

### 4.3.3 REPRESENTATION OF RATE-CHANGED UNVOICED SPEECH

If $x(n)$ now represents samples of an unvoiced-speech signal, then $x^{\beta}(n)$ is modelled as the output of the filter $t^{\beta}(n,m) = t(\beta n, m)$ driven by the white-noise process $u(n)$. Consequently, if the original speech $x(n)$ is characterized by its time-varying power spectrum (4.19) and autocorrelation function (4.18), then the rate-changed speech is characterized by the time-varying power spectrum

$$\begin{aligned} S_x^{\beta}(n,\omega) &= \sigma_u^2 |T_2(\beta n, \omega)|^2 \\ &= S_x(\beta n, \omega) \end{aligned} \tag{4.29}$$

and autocorrelation function

$$R_x^{\beta}(n,\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x^{\beta}(n,\omega)\exp[j\omega\tau]\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\beta n,\omega)\exp[j\omega\tau]\,d\omega$$

$$= R_x(\beta n,\tau)\ . \qquad (4.30)$$

Thus, changing the rate of an unvoiced-speech signal preserves the local statistics of the signal, but linearly time scales the time-varying parameters of the statistics.

# CHAPTER 5

# THEORY OF TIME-SCALE MODIFICATION OF SPEECH BASED ON SHORT-TIME FOURIER ANALYSIS

## 5.1 INTRODUCTION

The representation of a rate-changed speech signal, in terms of a model for the production of the original speech was formulated in Chapter 4. In order to actually change the rate of a given speech signal, the parameters of the representation must be estimated, appropriately modified, and used to synthesize the rate-changed signal. This chapter discusses short-time Fourier analysis as a method of accomplishing this objective.

The approach will be to modify the short-time Fourier transform of the given speech signal such that the signal synthesized from the modified transform corresponds to the desired rate-changed speech. The modification will be formulated, first, for voiced speech. Then, the same modification will be shown to produce rate changes of unvoiced speech as well.

Consequently, the problem of distinguishing voiced from unvoiced speech is avoided.

Heuristically, the use of short-time Fourier analysis for time-scale modification of speech is based on the idea of mapping the one-dimensional speech signal to a two-dimensional signal that is a function of time and frequency such that "temporal features" of the speech appear as functions of the time variable and "spectral features" appear as functions of the frequency variable. The short-time Fourier transform is then appropriately modified such that it is compressed or expanded along the time axis but not the frequency axis. This modification will require decimation and interpolation of the short-time Fourier transform in the time direction and a nonlinear modification, affecting its phase.

The distinction between temporal and spectral features of an audio signal is, in general, a fuzzy notion, based on the nature of the signal processing performed by the auditory system. Here, the distinction between temporal and spectral features for the class of speech signals is defined, based on the speech-production model and signal representation of Chapter 4. Specifically, the dependence on n of both the time-varying frequency response $T_2(n,\omega)$ and the pitch $\Omega(n)$, in the speech-production model, is assumed to represent the temporal characteristics of the speech, and the dependence on $\omega$ of $T_2(n,\omega)$ and the value of $\Omega(n)$ are assumed to represent the spectral characteristics of the speech. Unfortunately, the functions $T_2(n,\omega)$ and $\Omega(n)$ cannot, in general, be exactly determined by observing the speech waveform.

The technique of short-time Fourier analysis provides a means for estimating and modifying these functions. Expressing the short-time Fourier transform of $x(n)$ as the convolution in time:

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\,x(m)\exp[-j\omega m] \tag{5.1}$$

suggests that features of $x(n)$ that change slowly as a function of time over the duration of $h(n)$ appear in the short-time Fourier transform as functions of the time variable, whereas, features that change rapidly as a function of time over the duration of $h(n)$ appear as a function of the frequency variable. Equivalently, expressing the short-time Fourier transform, according to eqn. (2.30b), as the convolution in frequency:

$$X_2(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\psi)\,H(\psi-\omega)\exp[j(\psi-\omega)n]\,d\psi \tag{5.2}$$

suggests that features of $X(\omega)$ that change rapidly as a function of frequency over the width of $H(\omega)$ appear in the short-time Fourier transform as functions of the time variable, whereas, features of $X(\omega)$ that change slowly as a function of frequency appear as functions of the frequency variable.

## 5.2 SHORT-TIME FOURIER ANALYSIS OF VOICED SPEECH

According to the discussion of Chapter 4, a voiced-speech signal is modelled, on a short-time basis, as the sum of harmonically-related complex exponentials, each with slowly-varying complex amplitude, and instantaneous frequency. These harmonic components are each narrow-band signals, occupying non-overlapping frequency bands. If the short-time Fourier transform is interpreted as the output of a filter-bank spectrum analyzer, and if the analysis filter, $h(n)$, is designed such that its bandwidth is sufficiently wide to pass any one of the individual harmonic components of the speech, yet narrow enough to pass at most one such component, then short-time Fourier analysis provides a means of separating the individual harmonic components of the voiced speech signal. Once these harmonics have been separated, the individual components can be modified and used to synthesize rate-changed speech.

The short-time Fourier transform of a voiced-speech signal, $x(n)$, can be expressed in terms of the parameters of the harmonic representation for $x(n)$. Substituting the harmonic representation (4.15) for $x(n)$ into eqn. (5.1) gives

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{P(m)-1} h(n-m)\, c_k(m) \exp[jk\phi(m)] \exp[-j\omega m] \quad . \qquad (5.3)$$

Assuming the unit-sample response of the analysis filter, $h(n)$, is sufficiently short that the pitch of the speech is constant over the duration of $h(n)$, the local representation for $\phi(n)$ can be used to simplify eqn. (5.3). Specifically, $P(m)$ is replaced with $P(n)$, and the

representation

$$\phi(m) \approx \phi(n) + \Omega(n)(m-n) \qquad \text{for} \qquad h(n-m) \neq 0 \tag{5.4}$$

is substituted for $\phi(m)$ in eqn. (5.3) to give

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{P(n)-1} h(n-m) \times c_k(m) \exp[jk(\phi(n) + \Omega(n)(m-n))] \exp[-j\omega m]$$

$$= \sum_{k=0}^{P(n)-1} \sum_{m=-\infty}^{\infty} h(n-m)\, c_k(m) \exp[-j(\omega - k\Omega(n))m] \times \exp[jk(\phi(n) - \Omega(n)n)] \; . \tag{5.5}$$

The summation over m in eqn. (5.5) is recognized as the short-time Fourier transform of $c_k(n)$, evaluated for $\omega$ replaced by $[\omega - k\Omega(n)]$. According to the model for voiced speech, discussed in Chapter 4, the harmonic amplitudes, $c_k(n)$ are narrow-band low-pass sequences. If the analysis filter, $H(\omega)$, is designed with a bandwidth greater than the bandwidth of $C_k(\omega)$, the Fourier transform of $c_k(n)$, then the technique for evaluating the short-time Fourier transform of a narrow-band signal [Section 2.5] can be applied, for each value of k, to give

$$X_2(n,\omega) = \sum_{k=0}^{P(n)-1} c_k(n) H(-(\omega - k\Omega(n))) \exp[-j(\omega - k\Omega(n))n] \times \exp[jk(\phi(n) - \Omega(n)n)]$$

or,

$$X_2(n,\omega) = \sum_{k=0}^{P(n)-1} c_k(n) H(k\Omega(n)-\omega) \exp[j(k\phi(n)-\omega n)] \quad . \tag{5.6}$$

For the fixed value $n = n_o$, eqn. (5.6) expresses the short-time Fourier transform of $x(n)$ as the sum of $P(n_o)$ images of $H(\omega)$ each shifted in frequency by $k\Omega(n_o)$ and weighted by $c_k(n_o)\exp[j(k\phi(n_o)-\omega n_o)]$, as shown in Figure (5.1).

If the bandwidth of $H(\omega)$, in addition to being greater than the bandwidth of each of the $C_k(\omega)$'s, is also less than the instantaneous fundamental frequency, $\Omega(n)$, for all $n$, then the shifted and weighted images of $H(\omega)$ shown in Figure 5.1 are nonoverlapping as illustrated. $X_2(n,\omega)$, given by eqn. (5.6), therefore, reduces to

$$X_2(n,\omega) = \begin{cases} c_k(n) H(k\Omega(n)-\omega) \exp[j(k\phi(n)-\omega n)] & \\ \qquad\qquad \text{for } |\omega - k\Omega(n)| < \omega_h & \\ 0 \qquad\qquad \text{otherwise} & \end{cases} \tag{5.7}$$

where $\omega_h$ denotes the cutoff frequency of $H(\omega)$.

The short-time Fourier transform is a complex quantity, which can be expressed in polar form as

$$X_2(n,\omega) = A(n,\omega)\exp[j\Theta(n,\omega)] \tag{5.8}$$

where

$$A(n,\omega) = |X_2(n,\omega)|$$

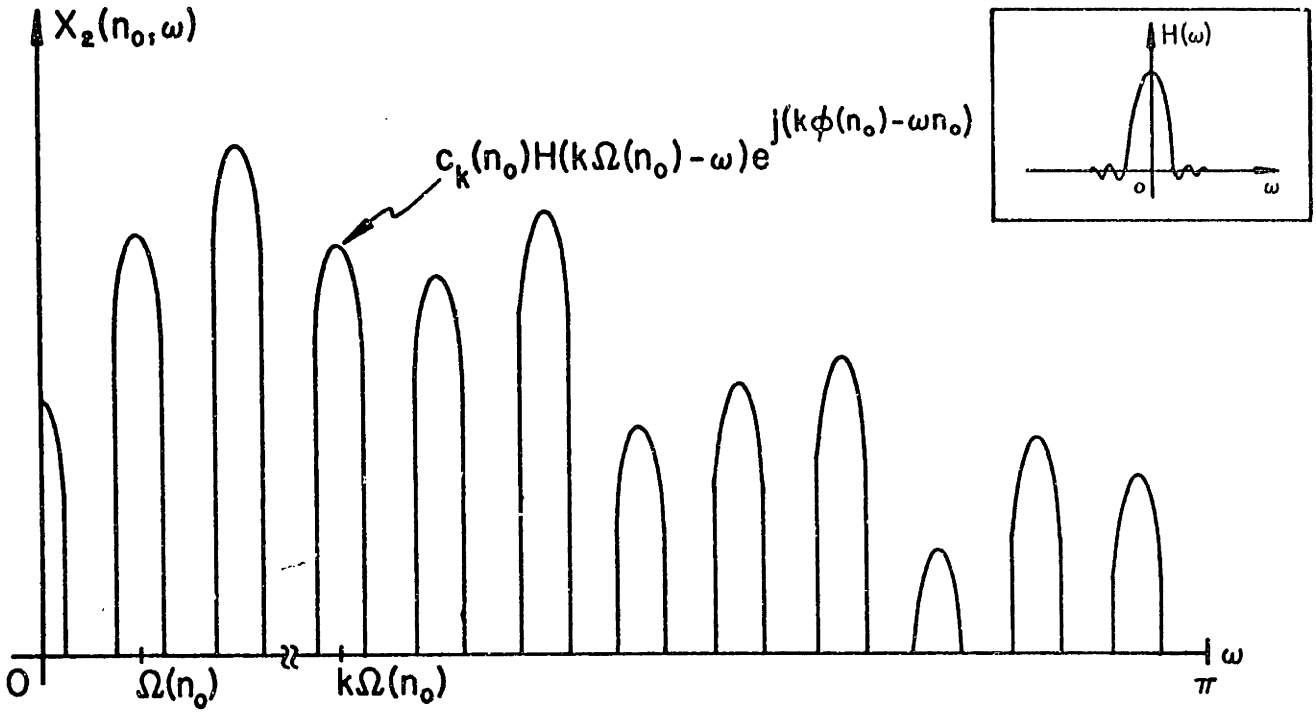



Figure 5.1

Short-Time Fourier Transform of an Idealized Speech Signal
Evaluated for the Particular Value of $n=n_0$

and

$$\Theta(n,\omega) = \arg[X_2(n,\omega)] \ .$$

Investigating the structure of $A(n,\omega)$ and $\Theta(n,\omega)$ will suggest how to modify $X_2(n,\omega)$ in order to obtain rate-changed speech. The magnitude of the short-time Fourier transform, obtained from eqn. (5.7) as

$$A(n,\omega) = |c_k(n)||H(k\Omega(n)-\omega)| \quad \text{for } |\omega-k\Omega(n)| < \omega_h \ , \qquad (5.9)$$

is a slowly-varying function of n because both $c_k(n)$ and $\Omega(n)$ are slowly-varying functions of n. Referring to eqn. (5.7), the phase of the short-time Fourier transform can be expressed as the sum of two components:

$$\Theta(n,\omega) = \alpha(n,\omega) + \beta(n,\omega) \qquad (5.10a)$$

where

$$\alpha(n,\omega) = \arg[c_k(n)] + \arg[H(k\Omega(n)-\omega)] \qquad (5.10b)$$

and

$$\beta(n,\omega) = k\phi(n) - \omega n \qquad (5.10c)$$

for

$$|\omega-k\Omega(n)| < \omega_h \ .$$

The component $\alpha(n,\omega)$ contributes a slowly time-varying phase and will be called the "phase-modulation" component. The other component, $\beta(n,\omega)$, can be expressed, using the definition (4.15c) of $\phi(n)$, as

$$\phi(n,\omega) = \begin{cases} \sum_{r=1}^{n} \{ k\Omega(r) - \omega \} & \text{for } n > 0 \\ 0 & \text{for } n = 0 \\ \sum_{r=-1}^{n} -\{ k\Omega(r) - \omega \} & \text{for } n < 0 \end{cases}$$

$$\text{where } |\omega - k\Omega(n)| < \omega_h . \tag{5.11}$$

Consequently, $\phi(n,\omega)$ satisfies the recursion relation

$$\phi(n,\omega) = \phi(n-1,\omega) + k\Omega(n) - \omega . \tag{5.12}$$

or,

$$\phi(n,\omega) = \phi(n-1,\omega) + \Omega(n,\omega) \tag{5.13}$$

where

$$\Omega(n,\omega) = k\Omega(n) - \omega \tag{5.14}$$

with k such that

$$|\Omega(n,\omega)| < \omega_h . \tag{5.15}$$

Because

$$\Omega(n,\omega) = \phi(n,\omega) - \phi(n-1,\omega), \tag{5.16}$$

$\Omega(n,\omega)$ will be referred to as the "instantaneous frequency" of the short-time Fourier transform. Furthermore, because $\Omega(n)$ is a slowly-varying function of n, $\Omega(n,\omega)$ is also a slowly-varying function of

n. Thus, $\phi(n,\omega)$ can be expressed locally as

$$\phi(n_0+\tau,\omega) \approx \phi(n_0,\omega) + \Omega(n_0,\omega)\tau \qquad (5.17)$$

for $(n_0+\tau)$ in the neighborhood about $n_0$ for which the speech is modelled as periodic. $\phi(n,\omega)$ will, therefore, be referred to as the "linear-phase," or "frequency-modulation," component of the short-time Fourier transform.

## 5.3 SYNTHESIS OF RATE-CHANGED VOICED SPEECH

The previous section provides the framework necessary to define the modified short-time Fourier transform from which rate-changed speech can be synthesized. Let $x(n)$ denote a given voiced-speech signal, and let $Y_2(n,\omega)$ denote the modified short-time Fourier transform from which the rate-changed speech, $x^{\beta}(n) = y(n)$, is to be synthesized. It will be shown that $Y_2(n,\omega)$ is given by

$$Y_2(n,\omega) = A(\beta n,\omega)\exp[j(\alpha(\beta n,\omega) + \phi(\beta n,\omega)/\beta)] , \qquad (5.18)$$

i.e., both the magnitude and phase of the short-time Fourier transform are linearly time scaled by $\beta$, and the frequency-modulation component of the phase is divided by $\beta$. Remember that $\phi(n,\omega)$ is the "unwrapped" phase, given by eqn. (5.11), and not its principal value. This distinction is important whenever $1/\beta$ is not an integer.

To show that $y(n)$ is the desired rate-changed speech, $x^{\beta}(n)$, the definitions (5.9) and (5.10) of the magnitude and phase of the short-time Fourier transform are substituted into the modified short-time Fourier transform (5.18) to obtain

$$Y_2(n,\omega) = \begin{cases} c_k(\beta n)\, H(k\Omega(\beta n) - \omega) \exp[j(k\phi(\beta n)/\beta - \omega n)] \\ \qquad \text{for } |\omega - k\Omega(\beta n)| < \omega_h \\ 0 \end{cases} \tag{5.19}$$

Because the shifted and weighted images of $H(\omega)$ are assumed to be nonoverlapping, $Y_2(n,\omega)$ can be written as the sum of these images:

$$Y_2(n,\omega) = \sum_{k=0}^{P(\beta n)-1} c_k(\beta n)\, H(k\Omega(\beta n) - \omega) \exp[j(k\phi(\beta n)/\beta - \omega n)] \,. \tag{5.20}$$

The synthesized signal, $y(n)$, is generated according to the short-time Fourier synthesis formula

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n-r)\, Y_2(r,\omega) \exp[j\omega n]\, d\omega \,. \tag{5.21}$$

Thus, substituting $Y_2(n,\omega)$, given by eqn. (5.20), into eqn. (5.21) gives

$$y(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,c_k(\beta r)\,H(k\Omega(\beta r)-\omega) \times \exp[j\{k\phi(\beta r)/\beta - \omega r\}]\exp[j\omega n]\,d\omega$$

$$= \sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,c_k(\beta r)\exp[jk\phi(\beta r)/\beta] \times \frac{1}{2\pi}\int_{-\pi}^{\pi} H(k\Omega(\beta r)-\omega)\exp[j\omega(n-r)]\,d\omega\ .$$

Integrating over $\omega$,

$$y(n) = \sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,h(r-n)\,c_k(\beta r) \times \exp[jk\{\phi(\beta r)/\beta + \Omega(\beta r)(n-r)\}]$$

and using the local representation (4.25) for $\phi(\beta n)/\beta$ gives

$$y(n) = \sum_{r=-\infty}^{\infty}\sum_{k=0}^{P(\beta r)-1} f(n-r)\,h(r-n)\,c_k(\beta r)\exp[jk\phi(\beta n)/\beta]\ .$$

Assuming that the time-scaled pitch, $P(\beta n)$, is constant over any interval less than the duration of $\{f(n)h(-n)\}$ so that,

$$P(\beta(n-\tau)) \approx P(\beta n) \qquad \text{for } f(\tau)h(-\tau) \neq 0$$

or equivalently (with $r = n-\tau$),

$$P(\beta r) \approx P(\beta n) \qquad \text{for } f(n-r)h(r-n) \neq 0$$

gives

$$y(n) = \sum_{k=0}^{P(\beta n)-1} \sum_{r=-\infty}^{\infty} f(n-r)\, h(r-n)\, c_k(\beta r) \exp[jk\phi(\beta n)/\beta] \ . \tag{5.22}$$

The summation over r in eqn. (5.22) is just the convolution of the time-scaled harmonic amplitudes, $c_k(\beta n)$, with the filter $(f(n)h(-n))$. The frequency response of this composite filter,

$$\sum_{n=-\infty}^{\infty} f(n)\, h(-n) \exp[-j\omega n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\omega+\psi) H(\psi)\, d\psi \ ,$$

has a bandwidth on the order of the sum of the bandwidths of $F(\omega)$ and $H(\omega)$. If this bandwidth is wider than the bandwidth of each of the $c_k(\beta n)$'s, then the $c_k(\beta n)$'s are passed by the composite filter with negligible distortion. Consequently,

$$y(n) = \sum_{k=0}^{P(\beta n)-1} c_k(\beta n) \exp[jk\phi(\beta n)/\beta] \ , \tag{5.23}$$

which is the desired rate-transformed speech signal, $x^{\beta}(n)$, given by eqn. (4.28).

To conclude this section, the assumptions used in formulating the preceding theory of time-scale modification of voiced speech based on short-time Fourier analysis will now be summarized. First, for the analysis of voiced speech, the analysis window, h(n), is assumed to be sufficiently narrow in time to resolve the temporal features of the speech and also sufficiently narrow in frequency to resolve the spectral features of the speech. On the one hand, to resolve temporal changes in the pitch, the duration of h(n) is assumed to be short enough that the pitch can be

regarded as constant over any time interval with duration less than that of $h(n)$. Further, to resolve temporal changes in the speech due to changes in the vocal-tract geometry, the bandwidth of $H(\omega)$ is assumed to be much greater than each of the bandwidths of the individual speech harmonics, $C_k(\omega)$. This assumption is roughly equivalent to the assumption that the harmonic amplitudes $c_k(n)$ can be regarded as constant over any interval shorter than the duration of $h(n)$. On the other hand, to resolve spectral features of the speech, i.e., the value of the pitch and the shape of the spectral envelope, the bandwidth of $H(\omega)$ must be less than the value of the pitch, $\Omega(n)$.

For the synthesis of rate-changed speech, the composite window $f(n)h(-n)$ is assumed to be sufficiently narrow in time to resolve the temporal features of the rate-changed speech. Specifically, the duration of the composite window $f(n)h(-n)$ is assumed to be short enough that the time-scaled pitch, $\Omega(\beta n)$, can be regarded as constant over any interval with duration less than $f(n)h(-n)$ and, the bandwidth of $f(n)h(-n)$ is assumed to be much greater than each of the bandwidths of the time-scaled harmonic amplitudes $c_k(\beta n)$.

## 5.4 SHORT-TIME FOURIER ANALYSIS OF UNVOICED SPEECH

Unvoiced speech is modelled as a quasi-stationary random process characterized by its second moments. Recalling the interpretation of the short-time Fourier transform as the output of a demodulator followed by a low-pass filter [Figure 2.2], the short-time Fourier transform of unvoiced speech can be interpreted as a set (indexed on $\omega$) of stochastic time series (in n). This section presents the relationships between the (second order) statistics of the short-time Fourier transform and the statistics of the analyzed and synthesized time sequences. The algebra involved in the calculations of these results is quite tedious, but, for the most part, straight forward. Therefore, the results will be presented here with their derivations deferred to the appendix.

### 5.4.1 THE SHORT-TIME FOURIER TRANSFORM OF UNVOICED SPEECH

Suppose, for the moment, that x(n) is a stationary random process. It can be shown that the short-time Fourier transform of x(n) is also a stationary random process for each $\omega$, i.e.,

$$E\{ X_2(n+\tau,\omega) X_2^*(n,\omega)\} \text{ is independent of } n.$$

Unfortunately, the time series corresponding to the short-time Fourier transform of x(n) obtained for two different values of $\omega$ are not jointly stationary, i.e.,

$E\{ X_2(n+\tau,\omega_1) X_2^*(n,\omega_2)\}$ depends on n for $\omega_1 \neq \omega_2$ .

The short-time Fourier transform can be expressed as

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m) \exp[-j\omega m]$$

$$= \Big( \sum_{m=-\infty}^{\infty} x(m)\, h(n-m) \exp[j\omega(n-m)] \Big) \exp[-j\omega n]$$

$$= \{ x(n) *_n (h(n) \exp[j\omega n]) \} \exp[-j\omega n]$$

and, therefore, can be interpreted as the response to x(n) of the linear time-invariant band-pass filter $h(n)\exp[j\omega n]$, multiplied (demodulated) by $\exp[-j\omega n]$. Because the responses of two linear time-invariant systems to the same stationary input are jointly stationary, the two random processes

$$X_2(n,\omega_1) \exp[j\omega_1 n]$$

and

$$X_2(n,\omega_2) \exp[j\omega_2 n]$$

are jointly stationary. Thus,

$$E\{ (X_2(n+\tau,\omega_1) \exp[j\omega_1 (n+\tau)])\, (X_2(n,\omega_2) \exp[j\omega_2 n])^* \}$$
$$= E\{ X_2(n+\tau,\omega_1) X_2^*(n,\omega_2) \} \exp[j\omega_1 \tau] \exp[j(\omega_1-\omega_2) n] \qquad (5.24)$$

is independent of n and, $X_2(n,\omega_1)$ and $X_2(n,\omega_2)$ are jointly stationary "to within a modulation by $\exp[j(\omega_1-\omega_2)n]$." Letting $\omega_1 = \omega - \epsilon/2$ and $\omega_2 = \omega + \epsilon/2$, the right-hand side of eqn. (5.24) becomes

$$\mathbb{E}\{ X_2(n+\tau,\omega_1) X_2^*(n,\omega_2)\} \exp[j\omega_1\tau] \exp[j(\omega_1-\omega_2)n]$$
$$= \mathbb{E}\{ X_2(n+\tau,\omega-\tfrac{\epsilon}{2}) X_2^*(n,\omega+\tfrac{\epsilon}{2})\} \exp[j(\omega-\tfrac{\epsilon}{2})\tau] \exp[-j\epsilon n]$$
$$= \mathbb{E}\{ X_2(n+\tau,\omega-\tfrac{\epsilon}{2}) X_2^*(n,\omega+\tfrac{\epsilon}{2})\} \exp[-j(n+\tfrac{\tau}{2})\epsilon] \exp[j\omega\tau] \qquad (5.25)$$

Because eqn. (5.25) is independent of n, a more convenient quantity than the cross-correlation function of the short-time Fourier transform is the modified correlation function, defined by

$$K_x(n,\omega,\tau,\epsilon) = \mathbb{E}\{ X_2(n+\tau,\omega-\tfrac{\epsilon}{2}) X_2^*(n,\omega+\tfrac{\epsilon}{2})\} \exp[-j(n+\tfrac{\tau}{2})\epsilon] \qquad (5.26)$$

which is independent of n. The cross-correlation function of the short-time Fourier transform can, therefore, be expressed as

$$\mathbb{E}\{ X_2(n+\tau,\omega-\tfrac{\epsilon}{2}) X_2^*(n,\omega+\tfrac{\epsilon}{2})\} = K_x(n,\omega,\tau,\epsilon) \exp[j(n+\tfrac{\tau}{2})\epsilon] \; . \qquad (5.27)$$

If x(n) is no longer stationary, but quasi-stationary, then the modified correlation function (5.26) becomes a "slowly-varying" function of n. By substituting the definition of the short-time Fourier transform (5.1) into the definition of the modified correlation function (5.26), the modified correlation function can be expressed in terms of the time-varying power spectrum of x(n) and the analysis filter $H(\omega)$ as

$$K_x(n,\omega,\tau,\epsilon) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n,\omega+\varphi) H(\varphi+\tfrac{\epsilon}{2}) H^*(\varphi-\tfrac{\epsilon}{2}) \exp[j\varphi\tau] \, d\varphi \; . \qquad (5.28)$$

Because the analysis filter, h(n), is chosen to be narrow in both time and frequency, $K_x(n,\omega,\tau,\epsilon)$ can be approximated by the first few terms of a two-dimensional power series in $\tau$ and $\epsilon$. This representation will be useful

for determining the power spectrum of unvoiced-speech synthesized from the non-linearly modified short-time Fourier transform (5.18), originally defined to effect rate changes of voiced speech. Define the moments up to second order of the analysis filter as

$$J_h = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\omega)|^2 d\omega = \sum_{n=-\infty}^{\infty} |h(n)|^2 \qquad (5.29a)$$

$$M_h = \frac{1}{2\pi J_h}\int_{-\pi}^{\pi} \omega |H(\omega)|^2 d\omega \qquad (5.29b)$$

$$m_h = \frac{1}{J_h}\sum_{n=-\infty}^{\infty} n|h(n)|^2 \qquad (5.29c)$$

$$D_h^2 = \frac{1}{2\pi J_h}\int_{-\pi}^{\pi} \omega^2 |h(\omega)|^2 d\omega \qquad (5.29d)$$

$$d_h^2 = \frac{1}{J_h}\sum_{n=-\infty}^{\infty} n^2 |h(n)|^2 \qquad (5.29e)$$

and

$$\mu_h = \frac{-j}{4\pi J_h}\int_{-\pi}^{\pi} \omega [\, H'(\omega) H^*(\omega) - H(\omega) H'^*(\omega)\,] d\omega \,, \qquad (5.29f)$$

Introducing a suitable time shift to $h(n)$ and frequency shift to $H(\omega)$ so that their first moments vanish, i.e.,

$$M_h = 0 \quad \text{and} \quad m_h = 0$$

eqn. (5.28) can be expanded in the two-dimensional power series

$$K_x(n,\omega,\tau,\epsilon) = J_x(n,\omega)\{1 - \frac{1}{2}[\, D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] + \ldots\} \quad (5.30)$$

where $J_x(n,\omega)$ is the smoothed spectrum

$$J_x(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega+\varphi)\,|H(\varphi)|^2\, d\varphi\ , \quad (5.31)$$

$D_h$ is the rms bandwidth of $H(\omega)$, $d_h$ is the rms duration of $h(n)$, and $\mu_h$ is a real number that vanishes if $h(n)$, or $H(\omega)$, or both are real.

Because the bandwidth of the analysis filter is narrow compared with the sampling frequency of the speech, $X_2(n,\omega)$ is a narrow-band low-pass random process in n for each value of $\omega$. Consequently, the short-time Fourier transform of unvoiced speech, expressed in polar form, exhibits slowly-varying amplitude and instantaneous frequency similar to the short-time Fourier transform of voiced speech. The major difference, however, is that unvoiced speech possesses no underlying harmonic structure, across the spectrum, as does voiced speech. The phase components $\alpha(n,\omega)$ and $\phi(n,\omega)$ for unvoiced speech will be defined as the values calculated by the same estimator used to to calculate these quantities for voiced speech (one such estimator will be discussed in the next chapter).

### 5.4.2 THE TIME-VARYING POWER SPECTRUM OF THE SYNTHESIZED SIGNAL

Let $Y_2(n,\omega)$ denote a short-time, or modified short-time Fourier transform and, let $y(n)$ denote the signal synthesized from $Y_2(n,\omega)$ according to the short-time Fourier synthesis formula (5.21). If $K_y(n,\omega,r,\epsilon)$ denotes the modified correlation function for $Y_2(n,\omega)$, and if the bandwidth of the synthesis filter is wider than the bandwidth of $K_y(n,\omega,r,\epsilon)$ in the n direction, then the time-varying power spectrum of $y(n)$ is

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \sum_{r=-\infty}^{\infty} K_y(n,\varphi,r,\epsilon)\exp[-j(\omega-\varphi)r]\,d\epsilon\,d\varphi \,. \tag{5.32}$$

## 5.5 SYNTHESIS OF RATE-CHANGED UNVOICED SPEECH

Suppose that the short-time Fourier transform of a given unvoiced speech signal, $x(n)$, has been expressed in the form

$$X_2(n,\omega) = A(n,\omega)\exp[j(\alpha(n,\omega) + \theta(n,\omega))] \tag{5.33}$$

where $A(n,\omega)$, $\alpha(n,\omega)$, and $\Omega(n,\omega) = \theta(n,\omega) - \theta(n-1,\omega)$ are slowly-varying functions of n. Define the modified short-time Fourier transform $Y_2(n,\omega)$, just as in the voiced-speech case, according to eqn. (5.18), by linearly time scaling the magnitude and phase of the short-time Fourier transform by $\beta$, and dividing the linear-phase component by $\beta$. This section argues that the time-varying power spectrum of the synthesized signal, $y(n)$, is approximately the the same as the time-varying power spectrum of the

desired rate-changed signal, $x^{\beta}(n)$.

The determination of the time-varying power spectrum of the synthesized signal, $y(n)$, is a non-linear stochastic-processes problem that, in general, has no closed form solution in terms of elementary functions. In order to make the problem more tractable and to gain some insight into the effects of the non-linear modification (5.18) of the short-time Fourier transform, several simplifying assumptions will be made. These assumptions are:

1. The underlying random process, $u(n)$, in the speech production model is Gaussian.
2. The spectral resolution of the filter-bank analyzer is sufficient to resolve the unvoiced-speech spectrum, $S_x(n,\omega)$, for the purpose of calculating the moments of $K_x(n,\omega,\tau,\epsilon)$. This assumption means, for example, that the average bandwidths and center frequencies of the filter-bank output signals are determined principally by the shape of the analysis filter, $H(\omega)$, rather than by fine structure in the spectrum $S_x(n,\omega)$. This assumption is reasonable because of the smoothness of the unvoiced-speech spectrum [Heinz and Stevens].
3. The slowly-varying phase component, $\alpha(n,\omega)$, of the short-time Fourier transform, will be neglected in the computation of the moments of $K_y(n,\omega)$. Therefore, the approximation

$$\begin{aligned} K_y(n,\omega,\tau,\epsilon) \approx \mathbb{E}\{ & A(\beta(n+\tau),\omega-\tfrac{\epsilon}{2})\, A(\beta n,\omega+\tfrac{\epsilon}{2}) \\ & \times \exp[j(\theta(\beta(n+\tau),\omega-\tfrac{\epsilon}{2}) - \theta(\beta n,\omega+\tfrac{\epsilon}{2}))/\beta]) \\ & \times \exp[-j(n+\tfrac{\tau}{2})\epsilon] \,. \end{aligned} \qquad (5.34)$$

will be used to calculate the moments of $K_y(n,\omega,\tau,\epsilon)$.

Based on these assumptions, the modified autocorrelation function (5.34) can be expanded in a power series in $\tau$ and $\epsilon$. The coefficients of this series are calculated (with considerable algebra) and expressed in terms of the moments of $K_x(n,\omega,\tau,\epsilon)$ to obtain

$$K_y(n,\omega,\tau,\epsilon) = J_x(\beta n,\omega)\left(1 - \frac{\gamma^2}{2}\left[D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2\right] + \ldots\right) \tag{5.35a}$$

where

$$\gamma^2 = \frac{1}{2}(1+\beta^2) \tag{5.35b}$$

Eqns. (5.35) again assume that the analysis filter has been defined with the appropriate shifts in time and frequency so that its first moments in time and frequency vanish. Comparing eqns. (5.30) and (5.35), shows that the modified correlation function for $Y_2(n,\omega)$ is given, to second order in $\tau$ and $\epsilon$, by

$$K_y(n,\omega,\tau,\epsilon) \approx K_x(\beta n,\omega,\gamma\tau,\gamma\epsilon/\beta) \tag{5.36}$$

The time-varying power spectrum of the synthesized rate-changed speech is given according to eqn. (5.32) as

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \sum_r K_x(\beta n,\varphi,\gamma r,\gamma\epsilon/\beta) \times \exp[-jr(\omega-\varphi)]\,d\epsilon\,d\varphi\,. \tag{5.37}$$

Now, substituting the expression (5.28) for $K_x(n,\omega,\tau,\epsilon)$ into the expression (5.37) gives

$$S_y(n,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\beta n, \omega+\varphi)\, G_\beta(\varphi)\, d\varphi \ . \qquad (5.38a)$$

where

$$G_\beta(\omega) = [\,\beta/\gamma(1-\gamma)\,]\cdot|F(\omega\gamma/(1-\gamma))|^2 H_1(\omega/(1-\gamma)) \qquad (5.38b)$$

and

$$H_1(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\omega+\tfrac{\theta}{2})\, H^*(\omega-\tfrac{\theta}{2})\, d\theta$$

$$= \frac{1}{\pi}\int_{-\pi}^{\pi} H(2\omega-\varphi)\, H^*(\varphi)\, d\varphi \qquad (5.38c)$$

Thus, $S_y(n,\omega)$ corresponds to the ideal spectrum of the rate-changed speech, smoothed by the function $G_\beta(\omega)$, whose width depends on the rate-change scale factor $\beta$.

The width of $G_\beta(\omega)$ is now shown to be sufficiently narrow that the spectrum, $S_y(n,\omega)$, for the synthesized speech is an acceptably close approximation to the ideal spectrum, $S_x(\beta n,\omega)$, for the rate-changed speech. The width of $G_\beta(\omega)$ is approximately the smaller of the widths of $|F(\omega\gamma/(1-\gamma))|^2$ and $H_1(\omega/(1-\gamma))$. In practice, the bandwidth of $F(\omega)$ is chosen to be approximately equal to $D_h$, the bandwidth of $H(\omega)$. Furthermore, from eqn. (5.38c), the bandwidth of $H_1(\omega)$ is seen to be approximately equal to $D_h$, also. Consequently, the width of the spectral smoothing function $G_\beta(\omega)$ is on the order of the smaller of $|(1-\gamma)/\gamma|D_h$ $|(1-\gamma)|D_h$.

Figure 5.2a shows plots of $|(1-\gamma)/\gamma|$ and $|1-\gamma|$ vs $\beta$ for $\beta \geq 1$, corresponding to time-scale compression. Here,

$$|(1-\gamma)/\gamma| \leq |1-\gamma|$$

and

$$|(1-\gamma)/\gamma| < 1.$$

Thus, for time-scale compression, the width of $G_\beta(\omega)$ is less than $|(1-\gamma)/\gamma|D_h$, which is less than $D_h$.

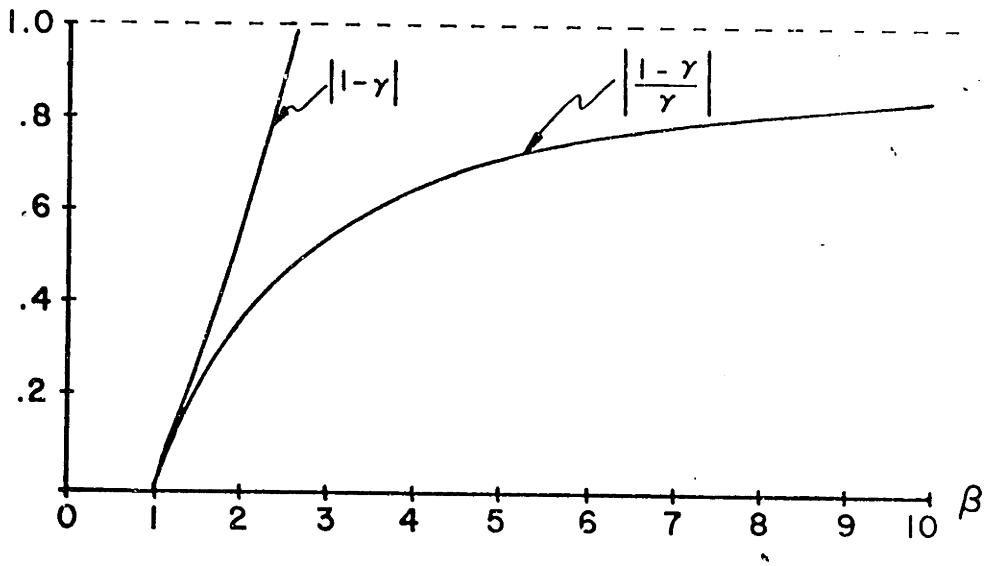
Figure 5.2b shows plots of $|(1-\gamma)/\gamma|$ and $|1-\gamma|$ vs $1/\beta$ for $\beta \leq 1$, corresponding to time-scale expansion. Here,

$$|1-\gamma| \leq |(1-\gamma)/\gamma|$$

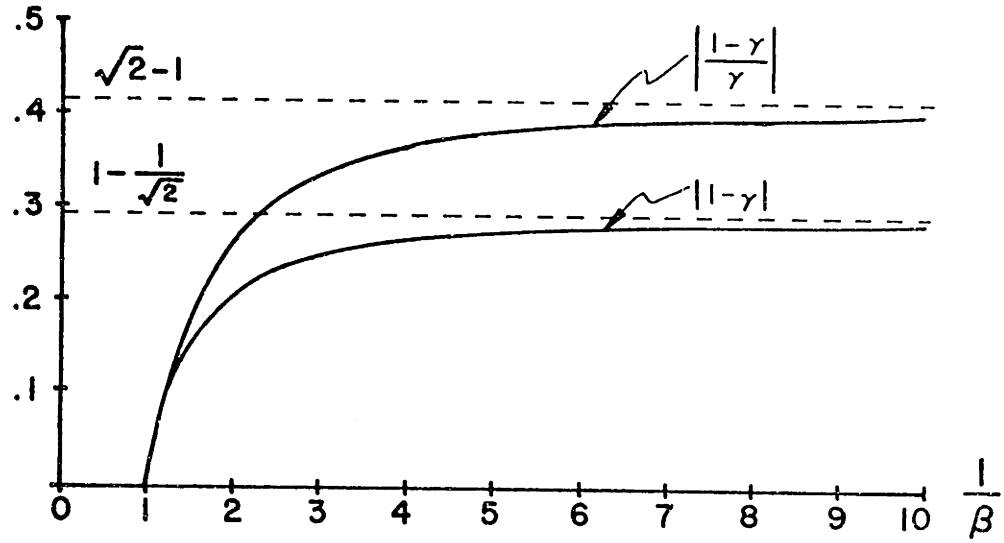
and

$$|1-\gamma| < 1 - 2^{-1/2} \approx 0.3 \quad .$$

Thus, for time-scale expansion, the width of $G_\beta(\omega)$ is less than $|1-\gamma|D_h$, which is less than $0.3D_h$. In practice, the bandwidth, $D_h$, of the analysis filter is on the order of 100 Hz. Thus, the spectral smearing of the synthesized speech is on the order of 100 Hz for time-scale compression and 30 Hz for time-scale expansion. The quality of real speech processed by a system based on this formulation confirms that the degree of spectral smearing of the unvoiced portions of the speech is acceptable.




(a)




(b)

Figure 5.2

Scale Factors for Width of Spectral-Smoothing Function for Rate-Changed Unvoiced Speech

# CHAPTER 6

# DESIGN AND SIMULATION OF A SYSTEM FOR TIME-SCALE MODIFICATION OF SPEECH BASED ON DISCRETE SHORT-TIME FOURIER ANALYSIS

## 6.1 INTRODUCTION

This chapter applies the theoretical results of the previous chapters to the design of a speech time-scale modification system which was simulated on a general purpose minicomputer. The system consists of an analysis / synthesis system which represents a given sampled speech signal in terms of its discrete short-time Fourier transform and a parameter modification system which appropriately modifies the short-time Fourier transform to effect the desired rate change in the synthesized speech. Furthermore, in the absence of any parameter modification the analysis / synthesis system is designed to be an identity system.

In order for the time-scale modification system to be realizable on a digital processor, the short-time Fourier transform must be represented by a finite number of frequency samples. Moreover, to make the amount of computation tractable, the short-time Fourier transform must be decimated in time (down sampled) as well. Consequently, in addition to the design of the analysis and synthesis filters based on the requirements of the previous chapters, a number of new issues arise as a result of the sampled-transform representation.

This chapter discusses the issues related to implementing a time-scale modification system for speech based on discrete short-time Fourier analysis and also introduces a procedure for estimating and modifying the fm component of the phase of the short-time Fourier transform. Section 6.2 raises the issue of how finely the short-time Fourier transform must be sampled to be adequately represented for the application of time-scale modification of speech. Section 6.3 discusses the analysis / synthesis system and, in particular, the considerations for designing the analysis and synthesis filters and choosing the temporal and spectral sampling intervals. Because the modification of the short-time Fourier transform used to effect rate changes of speech does not, in general, preserve the structure of the short-time Fourier transform, the results of Section 3.2 concerning the design of the analysis and synthesis filters and the sampling-rate requirements for the discrete short-time Fourier transform, cannot be applied directly. These issues must, therefore, be reconsidered for the particular problem of time-scale modification of speech. Section 6.4 discusses the modification of the short-time Fourier transform to

effect rate-changes in the synthesized speech. This modification consists of two operations: linearly time scaling the short-time Fourier transform by the rate-change scale factor $\beta$ and dividing the fm component of the phase of the short-time Fourier transform by the factor $\beta$. Because these two operations each affect the bandwidth of the short-time Fourier transform (considered as a time sequence), the order in which they are implemented becomes important to avoid aliasing of the sampled short-time Fourier transform. The linear time-scaling operation will, generally, be implemented using conventional digital interpolation / decimation techniques. Under certain circumstances, however, it may be computationally more efficient to effect the linear time scaling implicitly by performing the short-time Fourier analysis and synthesis assuming different temporal sampling rates. This issue will be discussed in Section 6.4.4. The phase modification will require removing jumps of $\pi$ and $2\pi$ in the phase curve (as a function of time) and introduces a further consideration in choosing the temporal sampling rate for the short-time Fourier transform; namely, that unwrapping the phase of the short-time Fourier transform requires samples of the principal value of the phase of the short-time Fourier transform at a rate of at least twice that of the Nyquist rate for the short-time Fourier transform. The chapter concludes with a discussion of the simulation on a Dec PDP-11/50 minicomputer.

## 6.2 TIME-SCALE MODIFICATION OF SPEECH BASED ON DISCRETE SHORT-TIME FOURIER ANALYSIS

Because the modification of the short-time Fourier transform required to effect rate changes of speech is non linear, the result of this modification applied to the discrete short-time Fourier transform is not, in general, equivalent to the result of this modification applied to the short-time Fourier transform. However, since the discrete short-time Fourier transform becomes the short-time Fourier transform, in the limit as the temporal and spectral sampling intervals approach unity and zero, respectively, the question arises: how fast must the discrete short-time Fourier transform be sampled in time and frequency so that the result of processing the samples of the short-time Fourier transform is "sufficiently close" to the result of processing the short-time Fourier transform itself? The non-linear nature of the processing makes this question difficult to answer in general. This section proposes an answer by obtaining conditions such that 1) with no parameter modification, the overall system is an identity system; 2) for the particular case of quasi-periodic voiced speech, the results of modifying the short-time Fourier transform and the discrete short-time Fourier transform are the same; and, 3) as the speech deviates from this model, either because the speech itself does not fit the quasi-periodic model, or because it has been corrupted by noise, the system is required to behave gracefully, i.e., the system is to be robust.

In order for the analysis / synthesis system to be an identity system, in the absence of parameter modification, the analysis and synthesis filters must satisfy the condition derived in Section 3.2. This condition

is that the effective filter $w(n,m)$, defined by

$$w(n,m) = \sum_{s=-\infty}^{\infty} f(n-sR)\,h(sR+m-n) \quad , \tag{6.1}$$

must have the property

$$w(n,pM) = 0 \qquad \text{for all } n, \tag{6.2}$$

where R is the temporal sampling interval, $\Omega_0 = 2\pi/M$ is the spectral sampling interval, and $h(n)$ and $f(n)$ are the analysis and synthesis filters.

For a voiced speech signal, $x(n)$, under quasi-stationary analysis, the modified short-time Fourier transform, $Y_2(n,\omega)$, given by eqn. (5.20), is the short-time Fourier transform of the desired rate-changed signal, $x^{\beta}(n)$, obtained using the same analysis filter, $h(n)$. Thus, for quasi-periodic voiced speech, condition (6.2) guarantees that the speech synthesized from the samples $Y_2(sR,k\Omega_0)$ will be the desired rate-changed speech $x^{\beta}(n)$.

If condition (6.2) is the sole criterion for designing the analysis /synthesis system, we find that in the absence of parameter modification, the simulated system is, indeed, an identity system, and for quasi-periodic speech (i.e., steady-state vowels) the rate-changed speech is high quality. As the speech deviates from the quasi-periodic model, however, an objectionable amount of reverberation becomes apparent in the rate-changed speech. This reverberation is, in fact, time-domain aliasing, resulting because the nonlinear operation required to effect the desired rate change of the speech is performed on the <u>samples</u> of the short-time Fourier transform, $X_2(sR,k\Omega_0)$, rather than on the short-time Fourier

transform, $X_2(n,\omega)$, itself. Thus, the reverberation can be reduced by increasing the number of frequency samples, M. In particular, if the effective window, $w(n,m)$, defined by eqn. (6.1) has finite duration in m, then the reverberation in the processed speech is effectively eliminated by choosing the number of frequency samples, M, to be equal to, or greater than this duration.

The modification of the short-time Fourier transform requires an estimate of certain parameters of the speech model imbedded in the short-time Fourier transform. In order to extract these parameters from the sampled short-time Fourier transform, $X_2(sR,k\Omega_0)$, the temporal sampling interval, R, must be small enough to prevent frequency aliasing of $X(\psi,k\Omega_0)$ in $\psi$. Since the analysis filter, h(n) is chosen to have low-pass characteristics, the interpretation of the short-time Fourier transform as the output of h(n) requires that the sampling frequency, $2\pi/R$, be greater than twice the cutoff frequency of $H(\omega)$ in order to prevent frequency aliasing.

## 6.3 DESIGN OF THE ANALYSIS / SYNTHESIS SYSTEM

The analysis /synthesis system represents a given speech signal in terms of its complex discrete short-time Fourier transform. Such a system, when applied to speech coding, is generally referred to as a phase vocoder [Flanagan and Golden]. With appropriate analysis and synthesis filters and proper sampling rates, the samples of the short-time Fourier transform are related to the parameters of the harmonic representation for voiced speech

and the time-varying spectral envelope of unvoiced speech as discussed in Chapter 4, thus, providing a tool for manipulating these parameters.

For a given input signal, $x(n)$ the short-time Fourier analyzer calculates samples of the short-time Fourier transform

$$X_2(sR, k\Omega_0) = \sum_{m=-\infty}^{\infty} h(sR-m)\, x(m) \exp[-j\Omega_0 km] \qquad (6.3)$$

where $\Omega_0 = 2\pi/M$ represents the spacing of the samples in frequency and R, the spacing of the samples in time. The analyzer calculates the samples $X_2(sR, k\Omega_0)$ efficiently with the FFT algorithm using the techniques discussed in Section (3.3.1).

The short-time Fourier synthesizer calculates samples of the time sequence, $y(n)$, from samples of the modified short-time Fourier transform, $Y_2(sR, k\Omega_0)$, according to the formula

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR)\, Y_2(sR, k\Omega_0) \exp[j\Omega_0 nk] \,. \qquad (6.4)$$

The synthesis is also implemented efficiently using the FFT algorithm as described in Section (3.3.2).

### 6.3.1 DESIGN OF THE ANALYSIS FILTER

As discussed in Chapter 4, the analysis filter, $h(n)$, must be sufficiently narrow band in frequency to resolve the "spectral features" of the speech, i.e., the bandwidth of $H(\omega)$ must be less than the spacing, in frequency, of the harmonics of the voiced portions of the speech, and small enough that the time-varying power spectrum of the unvoiced portions of the speech is smooth over any frequency interval smaller than this bandwidth. In addition, $H(\omega)$ must be sufficiently wide band to pass the temporal features of the speech. A final requirement on the analysis filter is dictated by the practical consideration of keeping the number of frequency samples small. Because the number of frequency samples, M, must be greater than the duration, in m, of $w(n,m)$, given by eqn. (6.1), to prevent time aliasing of the processed signal, $h(n)$ should have finite duration and be as short as possible.

In order to minimize the effects of time-domain aliasing in the discrete-transform implementation, a finite length Hamming window was chosen as the analysis filter $h(n)$. If N denotes the duration of $h(n)$, then $H(\omega)$ is approximately bandlimited to $4\pi/N$ [Oppenheim and Schafer], i.e., the width of the main lobe of $H(\omega)$ is approximately $8\pi/N$ and the peak amplitude of the side lobes is -41dB. If the samples $X_2(sR,k\Omega_0)$ are to represent $X_2(n,k\Omega_0)$ with no aliasing of $X(\psi,k\Omega_0)$ in $\psi$, then the temporal sampling interval, R, must be chosen less than N/4. Calculating $X_2(sR,k\Omega_0)$, therefore, corresponds to sliding $h(n)$ by no more than one fourth of its length for each FFT frame. In order to unwrap the phase of the short-time Fourier transform, however, it will be shown in Section

6.4.2 that it is necessary to sample at twice the Nyquist rate for $X_2(n, k\Omega_0)$. Thus, for the simulation, we choose

$$R \leq N/8 \qquad (6.5)$$

## 6.3.2 DESIGN OF THE SYNTHESIS FILTER

Because the non-linear modification does not, in general, preserve the structure of the short-time Fourier transform, the synthesized signal depends on the design of f(n). The interpretation of the discrete short-time Fourier transform as samples of the output of the bandlimited analysis filter, h(n), implies that $X_2(n, k\Omega_0)$ can be reconstructed from its samples $X_2(sR, k\Omega_0)$ by 1:R bandlimited interpolation. This result is valid for the modified short-time Fourier transform, as well, provided that the modification does not cause frequency aliasing of the sampled transform. The synthesis filter, f(n), is, therefore, chosen as a 1:R optimal bandlimited interpolating filter in order to guarantee proper interpolation of the speech parameters imbedded in the modified short-time Fourier transform.

A procedure that is particularly well suited to designing interpolating filters for the short-time Fourier synthesizer is the algorithm proposed by [Oetken et.al.] for designing optimal FIR digital interpolating filters. This procedure is attractive because it is a simple and efficient technique for designing filters of very high order. Furthermore, if the data to be interpolated is over sampled, then the design algorithm can exploit this property to improve the performance of

the filter.

An additional benefit of the choosing f(n) as an interpolating filter is that the duration of the effective filter w(n,m) is kept small. Since R must be less than the Nyquist interval for h(n), the design of f(n) as a 1:R bandlimited interpolating filter gives

$$w(n,m) = \sum_{s=-\infty}^{\infty} f(n-sR)\, h(sR+m-n)$$

$$= h(m) \qquad (6.6)$$

and the effective filter length is just the length of the analysis filter h(n). Other designs for f(n) will, in general, result in an effective filter length equal to the sum of the durations of f(n) and h(n). Since the number of frequency samples, M, required to prevent time aliasing of the synthesized signal increases with the duration of w(n,m), filter designs other than a bandlimited interpolating filter will require a larger number of frequency samples to represent the short-time Fourier transform.

## 6.4 DESIGN OF THE PARAMETER MODIFICATION SYSTEM

The modification of the short-time Fourier transform used to effect rate changes of speech consists of two basic procedures: linear time scaling and phase modification. The details of these two procedures will be considered individually, then their combination and incorporation into the overall processing scheme will be considered.

### 6.4.1 LINEAR TIME SCALING

In order to effect a rate change, by the factor $\beta$, in the synthesized signal, the magnitude and phase of the short-time Fourier transform must be linearly time scaled by the factor $\beta$. Because the magnitude and phase are obtained by non-linear operations on the short-time Fourier transform, these quantities have greater bandwidths than the real and imaginary parts of the short-time Fourier transform. Consequently, the linear time scaling is implemented by applying bandlimited decimation / interpolation techniques [Schafer and Rabiner, 1973b] to the real and imaginary parts of the short-time Fourier transform. The interpolating filter, denoted $f_M(n)$, is again designed by Oetken's technique.

Assuming the rate-change scale factor $\beta$ is a rational number, expressed as

$$\beta = D/I \tag{6.7}$$

bandlimited decimation / interpolation amounts to the direct evaluation of the sum

$$X_2(\beta sR, k\Omega_0) = \sum_{r=-\infty}^{\infty} f_M(sD-rI)\, X_2(rR, k\Omega_0) \tag{6.8a}$$

for the short-time Fourier transform, and

$$Y_2(sR, k\Omega_0) = \sum_{r=-\infty}^{\infty} f_M(sD-rI)\, Y_2(rR/\beta, k\Omega_0) \tag{6.8b}$$

for the modified short-time Fourier transform. This procedure is efficient because $f_M(n)$ has finite duration and the sum need only be evaluated once

to compute each value of $X_2(\beta sR, k\Omega_0)$ or $Y_2(sR, k\Omega_0)$.

### 6.4.2 PHASE MODIFICATION

The second procedure in modifying the phase of the short-time Fourier transform is dividing the fm component of the phase by the factor $\beta$. Specifically, given samples of the short-time Fourier transform expressed, according to the discussion in Chapter 4, as

$$X_2(sR', k\Omega_0) = a(sR', k\Omega_0)\exp[j\phi(sR', k\Omega_0)] \qquad (6.9)$$

where $\phi(n,\omega)$ denotes the fm component of the phase of the short-time Fourier transform, $a(n,\omega)$ is a complex function which varies slowly in n, and $R' = R$ or $R' = \beta \cdot R$ depending upon whether the phase modification is performed before or after the linear time-scaling operation, we wish to calculate the samples

$$Y_2(sR'/\beta, k\Omega_0) = a(sR', k\Omega_0)\exp[j\phi(sR', k\Omega_0)/\beta] \qquad (6.10)$$

Implementing this modification requires first estimating samples of the unwrapped fm phase component, $\phi(n,\omega)$, from samples of the principal value of the phase of the short-time Fourier transform. If

$$X_2(n,\omega) = a(n,\omega)\exp[j\phi(n,\omega)]$$

$$= A(n,\omega)\exp[j(\alpha(n,\omega) + \phi(n,\omega))]$$

$$= A(n,\omega)\exp[j\theta(n,\omega)]$$

where

$$a(n,\omega) = A(n,\omega)\exp[j\alpha(n,\omega)]$$

and

$$A(n,\omega) = |a(n,\omega)| = |X_2(n,\omega)|$$

$$\alpha(n,\omega) = \arg[a(n,\omega)]$$

$$\theta(n,\omega) = \alpha(n,\omega) + \phi(n,\omega) = \arg[X_2(n,\omega)] , \qquad (6.11)$$

then the samples of the fm phase component, $\phi(n,\omega)$, are to be estimated from samples of the principal value of the phase, denoted $PV[\theta(n,\omega)]$. As a function of n, $PV[\theta(n,\omega)]$ contains jumps of $2\pi$ due to the principal-value operator, and jumps of $\pi$ due to sign changes in the real and imaginary parts of $a(n,\omega)$. Although the real and imaginary parts of $a(n,\omega)$ are slowly-varying functions of n, the phase, $\alpha(n,\omega)$ can jump by $\pi$ when the real or imaginary part of $a(n,\omega)$ changes sign. Except for these jumps of $\pi$, $\alpha(n,\omega)$ is also a slowly-varying function of n. In order to estimate $\phi(n,\omega)$ from $PV[\theta(n,\omega)]$, it will be convenient to define the phase functions $\theta_\pi(n,\omega)$ and $\alpha_\pi(n,\omega)$, corresponding to $\theta(n,\omega)$ and $\alpha(n,\omega)$ with jumps of integer multiples of $\pi$ removed. Thus,

$$\theta_\pi(n,\omega) = \alpha_\pi(n,\omega) + \phi(n,\omega) \qquad (6.12)$$

Once $\theta_\pi(n,\omega)$ has been determined, $\phi(n,\omega)$ is estimated from $\theta_\pi(n,\omega)$ using the property that $\alpha_\pi(n,\omega)$ and the first difference of $\phi(n,\omega)$ are both slowly-varying functions of n.

A procedure will now be developed for determining $\theta_{\pi}(n,\omega)$ from $PV[\theta(n,\omega)]$, based on the interpretation of the short-time Fourier transform as the output of the low-pass filter $h(n)$. Denoting the first backward difference operator, with respect to n, by $\nabla_n$, the first difference of the principal value of the phase of the short-time Fourier transform is

$$\begin{aligned}\nabla_n PV[\theta(n,\omega)] &= PV[\theta(n,\omega)] - PV[\theta(n-1,\omega)] \\ &= \nabla_n \theta_{\pi}(n,\omega) + \pi I_1(n,\omega) + 2\pi I_2(n,\omega)\end{aligned} \tag{6.13}$$

where $I_1$ and $I_2$ are integer-valued functions of n and $\omega$ representing the jumps of $\pi$ and $2\pi$ in $PV[\theta(n,\omega)]$. Therefore, $\nabla_n \theta_{\pi}(n,\omega)$ differs from $\nabla_n PV[\theta(n,\omega)]$ only by an unknown integer multiple of $\pi$. Since $\nabla_n \theta_{\pi}(n,\omega)$ is the instantaneous frequency of $X_2(n,\omega)$, and since $X_2(n,\omega)$ is the output of the (low-pass) analysis filter, $h(n)$, $|\nabla_n \theta_{\pi}(n,\omega)|$ is assumed to be less than the cutoff frequency of $h(n)$. If $\omega_h$ denotes the cutoff frequency of $h(n)$, then

$$|\nabla_n \theta_{\pi}(n,\omega)| < \omega_h . \tag{6.14}$$

Furthermore, if $\omega_h < \pi/2$, then

$$|\nabla_n \theta_{\pi}(n,\omega)| < \frac{\pi}{2} \tag{6.15}$$

and $\nabla_n \theta_{\pi}(n,\omega)$ can be determined from $\nabla_n PV[\theta(n,\omega)]$ simply by adding integer multiples of $\pi$ to $\nabla_n PV[\theta(n,\omega)]$ until the result satisfies condition (6.15). $\theta_{\pi}(n,\omega)$ can then be reconstructed by the running sum

$$\theta_{\pi}(n,\omega) \approx \sum_{r=n_0+1}^{n} \nabla_r \theta_{\pi}(r,\omega) + \theta_{\pi}(n_0,\omega) \tag{6.16}$$

where $n_0$ is an initial time at which $\theta_{\pi}(n_0,\omega)$ is assumed to be

$$\theta_{\pi}(n_0,\omega) = PV[\theta(n_0,\omega)]$$

For the sampled transform implementation,

$$\nabla_s \theta_{\pi}(sR',k\Omega_0) \approx R'\nabla_n \theta_{\pi}(n,k\Omega_0)\Big|_{n=sR'} \tag{6.17}$$

because $\nabla_n \theta(n,k\Omega_0)$ is a slowly-varying function of n. Multiplying both sides of (6.14) by R′ and substituting (6.17) gives

$$|\nabla_s \theta_{\pi}(sR',k\Omega_0)| < \omega_h R' \tag{6.18}$$

Now, if $\omega_h R' < \pi/2$, corresponding to choosing the sampling frequency, $2\pi/R'$ greater than $4\omega_h$ (twice the frequency required by the sampling theorem for sampling the output of h(n)), eqn. (6.13) and condition (6.15) become

$$\nabla_s PV[\theta(sR',k\Omega_0)] = \nabla_s \theta_{\pi}(sR',k\Omega_0) + \pi I_1'(s,k) + 2\pi I_2'(s,k) \tag{6.19}$$

and

$$|\nabla_s \theta_{\pi}(sR',k\Omega_0)| < \frac{\pi}{2} \tag{6.20}$$

respectively. $\nabla_s \theta_{\pi}(sR',k\Omega_0)$ can, therefore, be determined from $\nabla_s PV[\theta(sR',k\Omega_0)]$ by adding integer multiples of $\pi$ until (6.20) is satisfied.

The problem of estimating the fm component of the phase, $\phi(n,\omega)$, from $\theta_{\pi}(n,\omega)$ is basically a problem of curve fitting. Since

$$\theta_{\pi}(n,\omega) = \alpha_{\pi}(n,\omega) + \phi(n,\omega) \ , \tag{6.21}$$

where $\alpha_{\pi}(n,\omega)$ is a slowly-varying function of n with

$$\nabla_n \alpha_{\pi}(n,\omega) \approx 0 \ ,$$

and $\phi(n,\omega)$ has a first difference,

$$\nabla_n \phi_{\pi}(n,\omega) = \Omega(n,\omega) \ ,$$

which is also a slowly-varying function of n, $\Omega(n,\omega)$ is the slope of a first-order polynomial (in n) that locally fits $\theta_{\pi}(n,\omega)$. While there are a variety of approaches to this problem, the technique to be described here was chosen for its simplicity and good performance in the actual simulation.

The first backward difference of eqn. (6.21) is given by

$$\begin{aligned} \nabla_n \theta_{\pi}(n,\omega) &= \nabla_n \alpha_{\pi}(n,\omega) + \nabla_n \phi(n,\omega) \\ &= \nabla_n \alpha_{\pi}(n,\omega) + \Omega(n,\omega) \\ &\approx \Omega(n,\omega) \end{aligned} \tag{6.22}$$

and the first forward difference by

$$\Delta_n \theta_{\pi}(n,\omega) = \theta_{\pi}(n+1,\omega) - \theta_{\pi}(n,\omega)$$

$$= \Delta_n \phi_{\pi}(n,\omega) + \Delta_n \vartheta(n,\omega)$$

$$= \Delta_n \phi_{\pi}(n,\omega) + \Omega(n+1,\omega) \qquad (6.23)$$

$$\approx \Omega(n,\omega)$$

A reasonable estimate for $\Omega(n,\omega)$, denoted $\tilde{\Omega}(n,\omega)$, is, therefore, the average of (6.22) and (6.23), given by

$$\tilde{\Omega}(n,\omega) = \frac{1}{2}[\Delta_n + \nabla_n]\ \theta_{\pi}(n,\omega)$$

$$= \frac{1}{2}[\theta_{\pi}(n+1,\omega) - \theta_{\pi}(n-1,\omega)]$$

$$= \mu\delta_n \theta_{\pi}(n,\omega)$$

where $\mu\delta_n = \frac{1}{2}[\Delta_n + \nabla_n]$ is known as the mean central-difference operator [Hildebrand]. In practice, since

$$\Delta_n \theta_{\pi}(n,\omega) = \nabla_n \theta_{\pi}(n+1,\omega)$$

$\theta_{\pi}(n,\omega)$ is never actually computed from (6.16), but $\tilde{\Omega}(n,\omega)$ is computed directly as the average of the forward and backward differences

$$\tilde{\Omega}(n,\omega) = \frac{1}{2}[\Delta_n + \nabla_n]\ \theta_{\pi}(n,\omega)$$

$$= \frac{1}{2}[\nabla_n \theta_{\pi}(n+1,\omega) + \nabla_n \theta_{\pi}(n,\omega)] \ . \qquad (6.24)$$

The estimate $\tilde{\vartheta}(n,\omega)$, for the fm phase component, is constructed from the running sum

$$\tilde{\vartheta}(n,\omega) = \sum_{r=1}^{n} \tilde{\Omega}(r,\omega) \tag{6.25}$$

where $\tilde{\vartheta}(0,\omega) = \vartheta(0,\omega) = 0$.

For the sampled transform implementation, the estimate for $\Omega(sR',k\Omega_0)$ becomes

$$\begin{aligned}\tilde{\Omega}(sR',k\Omega_0)R' &= \mu\delta_s\Theta_\pi(sR',k\Omega_0)\\ &= \tfrac{1}{2}[\Delta_s + \nabla_s]\,\Theta_\pi(sR',\omega)\\ &= \tfrac{1}{2}[\nabla_s\Theta_\pi(sR'+R',k\Omega_0) + \nabla_s\Theta_\pi(sR',k\Omega_0)]\end{aligned} \tag{6.26}$$

and

$$\tilde{\vartheta}(sR',k\Omega_0) = \sum_{r=1}^{s} \tilde{\Omega}(sR',k\Omega_0)R' \tag{6.27}$$

Once the estimate $\tilde{\vartheta}(sR',k\Omega_0)$ of the sampled fm phase component $\vartheta(sR',k\Omega_0)$ is calculated from eqn. (6.27), samples of the the phase-modified short-time Fourier transform are calculated by adding $(\frac{1}{\beta} - 1)\tilde{\vartheta}(sR',k\Omega_0)$ to the phase of $X_2(sR',\Omega_0)$, or equivalently, multiplying $X_2(sR',\Omega_0)$ by $\exp[j(\frac{1}{\beta} - 1)\tilde{\vartheta}(sR',k\Omega_0)]$, to obtain

$$\begin{aligned}Y_2(sR'/\beta,k\Omega_0) &\approx a(sR',k\Omega_0)\exp[j\vartheta(sR',k\Omega_0)]\exp[j(\tfrac{1}{\beta}-1)\tilde{\vartheta}(sR',k\Omega_0)]\\ &= a(sR',k\Omega_0)\exp[j(\vartheta(sR',k\Omega_0) + (\tfrac{1}{\beta}-1)\tilde{\vartheta}(sR',k\Omega_0))]\\ &\approx a(sR',k\Omega_0)\exp[j\vartheta(sR',k\Omega_0)/\beta]\ .\end{aligned} \tag{6.28}$$

### 6.4.3 THE OVERALL MODIFICATION SYSTEM

The linear time scaling and phase modification of the short-time Fourier transform each affect the bandwidth of $X_2(n,\omega)$, considered as a sequence in n for each $\omega$. If the short-time Fourier transform is down-sampled, close to its Nyquist rate, to obtain $X_2(sR,\omega)$, then the order in which the linear time scaling and phase modification are implemented becomes important to prevent frequency aliasing of $X(\psi,\omega)$ in $\psi$. In particular, the bandwidth (in $\psi$) of the time-scaled short-time Fourier transform,

$$X_2(\beta n,\omega) = a(\beta n,\omega)\exp[j\phi(\beta n,\omega)] ,$$

is $\beta$ times the bandwidth of the original short-time Fourier transform, $X_2(n,\omega)$. In contrast, the bandwidth (in $\psi$) of the phase-modified short-time Fourier transform,

$$Y_2(n/\beta,\omega) = a(n,\omega)\exp[j\phi(n,\omega)/\beta] ,$$

is approximately $1/\beta$ times the bandwidth of $X_2(n,\omega)$. The bandwidth of the modified short-time Fourier transform,

$$Y_2(n,\omega) = a(\beta n,\omega)\exp[j\phi(\beta n,\omega)/\beta] ,$$

obtained as the result of both linear time scaling and phase modification is approximately the same as the bandwidth of $X_2(n,\omega)$. Thus, for time-scale expansion ($0 < \beta < 1$), the linear time-scaling operation decreases the bandwidth of the short-time Fourier transform in $\psi$, while the phase modification operation increases it. Conversely, for time scale-compression

($\beta > 1$), the linear time-scaling operation increases the bandwidth of the short-time Fourier transform, while the phase modification decreases it.

Suppose $X_2(n,\omega)$ is represented by its samples $X_2(sR, k\Omega_0)$, where R is close to the Nyquist interval for sampling $X_2(n,\omega)$ in n. More precisely, suppose that either $\beta \cdot R$, or $\frac{1}{\beta} \cdot R$ is greater than the Nyquist interval for $X_2(n,\omega)$. Then, in order to prevent frequency aliasing (in $\psi$) when implementing time-scale expansion, the linear time-scaling must be implemented first, followed by the phase modification; conversely, when implementing time-scale compression, the phase modification must be implemented first, followed by the linear time scaling.

### 6.4.4 IMPLICIT TIME-SCALING

Under certain circumstances, computational savings may be gained by incorporating the linear time-scaling of the short-time Fourier transform into the analysis and synthesis procedures. This method will be called the implicit method for linearly time-scaling the short-time Fourier transform, in contrast to the previously described explicit method and is effected by assuming different temporal sampling rates for the short-time Fourier analysis and synthesis. Let $R_A$ denote the sampling interval at the output of the short-time Fourier analyzer, and $R_S$ the sampling interval assumed by the synthesizer. The output of the analyzer is, therefore, given by

$$X_2(sR_A, k\Omega_0) = \sum_{m=-\infty}^{\infty} h(sR_A - m)\, x(m) \exp[-j\Omega_0 km]$$

$$= a(sR_A, k\Omega_0) \exp[j\phi(sR_A, k\Omega_0)] \qquad (6.29)$$

and the output of the synthesizer given by

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR_S)\, Y_2(sR_S, k\Omega_0) \exp[j\Omega_0 nk] \; . \qquad (6.30)$$

If $Y_2(sR_S, k\Omega_0) = Y_2(sR_A/\beta, k\Omega_0)$ is defined as the result of dividing the fm component of the phase of short-time Fourier transform by $\beta$ (with no time-scaling), i.e.,

$$Y_2(sR_S, k\Omega_0) = a(sR_A, k\Omega_0) \exp[j\phi(sR_A, k\Omega_0)/\beta] \qquad (6.31)$$

then substituting eqn. (6.31) into eqn. (6.30) gives the expression for the synthesized signal

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n - sR_S)$$

$$\times\, a(sR_A, k\Omega_0) \exp[j\phi(sR_A, k\Omega_0)/\beta] \exp[j\Omega_0 nk] \; . \qquad (6.32)$$

Assuming $\beta = D/I$, let $R_A = R/I$ and $R_S = R/D$, where R is an integer parameter which specifies the actual sampling intervals such that $R_A$ and $R_S$ are integers. Thus, eqn. (6.32) becomes

$$y(n) = \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} f(n-sR/D) \times a(sR/I, k\Omega_0) \exp[j\phi(sR/I, k\Omega_0)/\beta] \exp[j\Omega_0 nk] \ . \tag{6.33}$$

If the sampling parameter R is chosen with careful attention to the issues discussed in the preceding sections, namely, that $X_2(sR_A, k\Omega_0) = X_2(sR/I, k\Omega_0)$ corresponds to samples of $X_2(n,\omega)$ sampled often enough to estimate the fm component of its phase, and $Y_2(sR_S, k\Omega_0) = Y_2(sR/I, k\Omega_0)$ is sampled often enough to prevent aliasing then, if $f(n)$ is a $1{:}R_S$ bandlimited interpolating filter, eqn. (6.33) becomes

$$\begin{aligned} y(n) &= \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} a(nD/I, k\Omega_0) \exp[j\phi(nD/I, k\Omega_0)/\beta] \exp[j\Omega_0 nk] \\ &= \frac{1}{M} \sum_{k=0}^{M-1} \sum_{s=-\infty}^{\infty} a(\beta n, k\Omega_0) \exp[j\phi(\beta n, k\Omega_0)/\beta] \exp[j\Omega_0 nk] \ , \end{aligned} \tag{6.34}$$

which is the desired rate-changed signal.

Whether or not the implicit method of linearly time-scaling the short-time Fourier transform is more efficient than the explicit method depends primarily on the factor $\beta$. For example, if either D or I is unity, then the implicit method is more efficient, whereas, if D and I are both large integers then the explicit method is generally more efficient. Another consideration in choosing between these two methods is the mode of operation of the system. For example, if the same speech passage is to be

processed several times with different rate-change scale factors, i.e., for perceptual studies, then it may be most efficient to perform the analysis once, at a relatively high sampling rate, store the samples of the short-time Fourier transform, and perform the synthesis assuming different sampling rates as described above. Another useful aspect of the implicit method is that it permits exploiting the requirement that the output of the short-time Fourier analyzer is sampled at twice the minimum rate required to avoid frequency aliasing. Thus, for example, time-scale expansion by the factor of 2:1 ($\beta = \frac{1}{2}$) can be implemented simply by modifying the phase of the discrete short-time Fourier transform and performing the synthesis assuming $R_S = 2 \cdot R_A$. Finally, the implicit and explicit methods are easily combined, in practice, to realize the advantages of both.

## 6.5 SIMULATION OF A TIME-SCALE MODIFICATION SYSTEM

The concepts of the previous sections were employed in a general purpose minicomputer simulation of a complete time-scale modification system for speech. This section discusses some of the details and results of the simulation.

The complete time-scale modification system is depicted in Figure 6.1. The figure is segmented into three sections: a short-time Fourier analyzer, a system for modifying the samples of the short-time Fourier transform, and a short-time Fourier synthesizer. Note that the two realizations of the modification system, one for time-scale expansion ($0 < \beta < 1$) and the other for time-scale compression ($\beta > 1$), are shown explicitly.

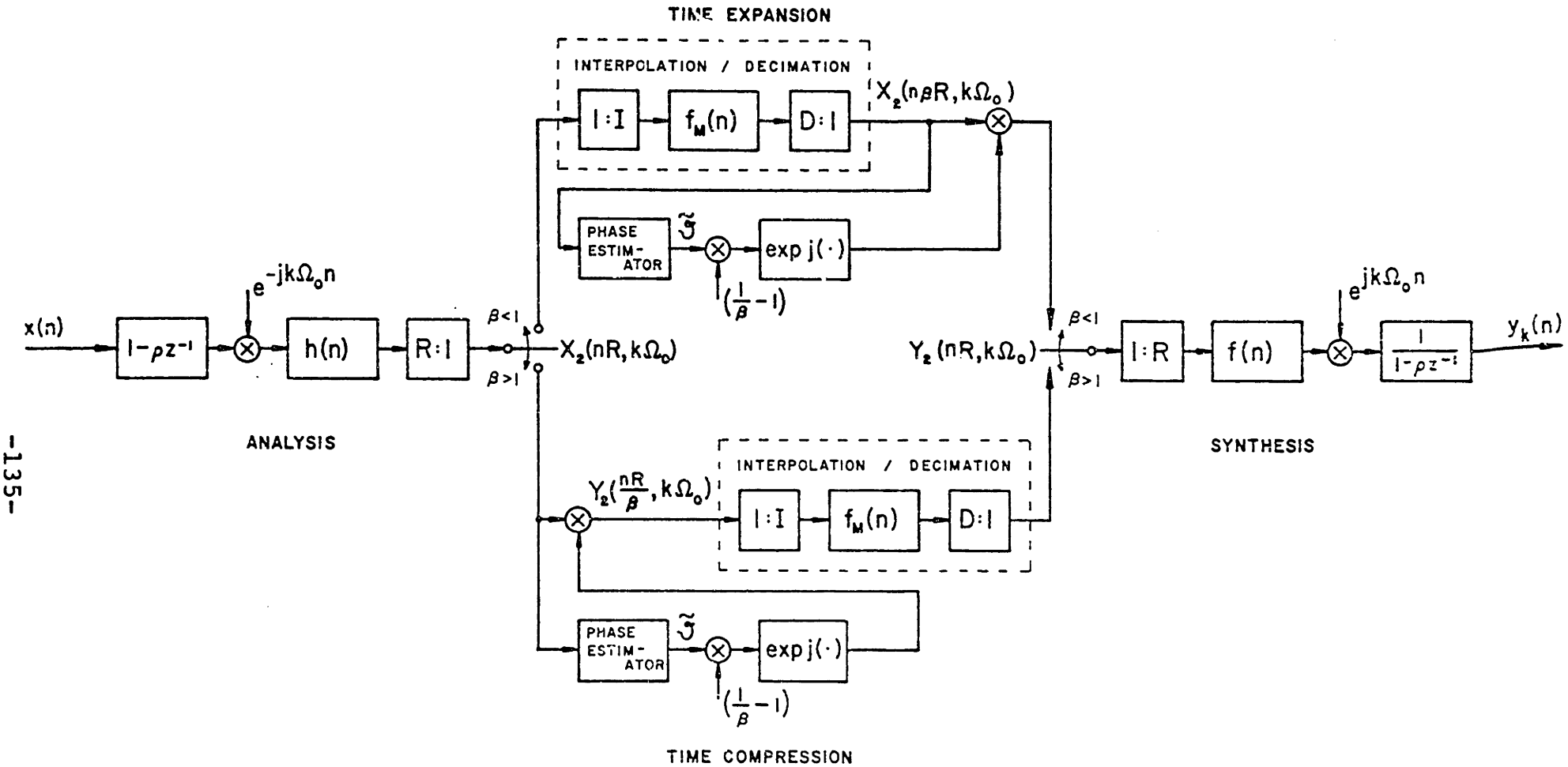



Figure 6.1

Block Diagram for One Channel of a Speech Rate-Change System

$$y(n)=\frac{1}{M}\sum_{k=0}^{M-1} y_k(n)$$

For the simulation, the input sequence, x(n), was obtained by sampling 5kHz low-pass filtered speech at the sampling rate of 10kHz. x(n) was then preemphasized using the first-order system

$$H_p(z) = 1 - \rho z^{-1} \tag{6.35}$$

with $\rho = 0.95$. The analysis window was chosen as an N-point Hamming window with N in the range $256 < N < 512$, corresponding to analog filter bandwidths of 156Hz to 78Hz. The output of the short-time Fourier analyzer was sampled every $R_A \leq \frac{N}{8}$ samples, and the number of frequency samples, M, was fixed at 512. The synthesis was performed assuming the sampling interval, $R_S$, less than N/4 and exploiting the oversampling of the transform, as described in Section (6.4.2). The synthesis filter was an optimal $1{:}R_S$ FIR digital interpolating filter designed using Oetken's method, taking into account the cutoff frequency of the data. For the explicit method of time-scaling the short-time Fourier transform, the interpolating filter $f_M(n)$ was designed with the same technique. The output sequence, y(n), was postemphasized with the system

$$H_p^{-1}(z) = 1/(1 - \rho z^{-1}) , \tag{6.36}$$

desampled, assuming the sampling rate of 10kHz, and low-pass filtered at 5kHz.

The results of the simulation demonstrate that this system is capable of producing high-quality rate-changed speech for reasonable values of $\beta$, i.e., for compression ratios as high as 3:1 and expansion ratios as high as

4:1 [Portnoff, 1977]. The processed speech retained its natural quality and speaker dependent features and was free from artifacts such as glitches, burbles, and reverberation, often present in vocoded speech. For time-scale expansion by greater than 4:1, the processed speech began to exhibit small amounts of reverberation, due to time-domain aliasing, which one should be able to eliminate by increasing the number of frequency samples, M. For time-scale compression of 4:1, or greater, the processed speech often became rough. Note, however, that while arbitrarily high expansion ratios are physically reasonable, arbitrarily high compression ratios are not. Consider, for example, a voiced phoneme containing four pitch periods. Greater than 4:1 compression reduces this phoneme to less than one pitch period, destroying the periodic character of the phoneme. Thus, one might expect speech, time compressed with high compression ratios, to have a rough quality and low intelligibility.

In addition to producing high-quality rate-changed speech when the original speech was high quality, the system also performed well processing degraded speech. Specifically, speech corrupted by additive Gaussian white noise with 0 dB signal-to-noise ratio was processed with good results [Portnoff, 1977].

# CHAPTER 7

# SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

## 7.1 SUMMARY

This dissertation develops certain aspects of the theory of short-time Fourier analysis, its application to speech, and in particular, its application to time-scale modification of speech. Short-time Fourier analysis provides both a time-frequency representation of signals and linear systems and an efficient technique for implementing a class of linear-filtering operations. When applied to speech, these concepts lead to the quasi-stationary representation developed in Chapter 4 and the short-time Fourier transform as a useful technique for speech processing. In addition, the theory provides both a mathematical framework for defining rate-changed speech and the basis for a system to effect the rate changes. Such a system was designed and implemented on a general-purpose minicomputer and was demonstrated to be capable of producing high-quality rate-changed speech.

## 7.2 SUGGESTIONS FOR FURTHER RESEARCH

Short-time Fourier analysis is a special case of the more general concept of representing a one-dimensional signal by a multi-dimensional signal. This concept may be useful simply because of computational considerations, e.g., [Agarwal and Burrus], but is especially useful if the representation maps different classes of features of the signal to different dimensions. For the case of short-time Fourier analysis, features characterized as slowly varying are mapped along the time axis, whereas, features characterized as rapidly varying are mapped along the frequency axis. This mapping is especially significant for acoustic signals because, in many respects, it is analogous to the processing of acoustic signals by the human auditory system [Gambardella, 1970; Callahan]. Perhaps, other such mappings of signals of one dimensionality to signals of a different dimensionality can be exploited in other classes of signal-processing problems. The treatment of short-time Fourier analysis in this thesis is restricted to short-time Fourier transforms defined with an analysis window that is independent of time and frequency. For some applications, this restriction may be undesirable. For example, a "constant-Q" representation may be desired for certain applications, [Gambardella, 1971; Youngberg and Boll] or, a data-adaptive analysis window desired for other applications [Wang; Patisaul and Hammett].

For the application of short-time Fourier analysis to rate changes of speech, a number of refinements and generalizations of the system described are possible. One such refinement is an improved estimate of the fm component of the phase of the short-time Fourier transform, perhaps, by a

better estimator than the simple difference scheme described in Chapter 6, or, perhaps, by coherent processing of adjacent frequency samples of the short-time Fourier transform. The major difficulty in designing a speech rate-change system based on short-time Fourier analysis results from the uncertainty principle, i.e., the analysis window cannot be arbitrarily short in time and arbitrarily narrow in frequency. Consequently, choosing the analysis window represents a compromise between the requirements of resolving the pitch of the speech (in the frequency domain), and resolving the temporal events of the speech (in the time domain). At times, the assumption that both of these requirements can be satisfied simultaneously is a borderline assumption. One possible approach to this problem is to improve the spectral estimate using coherent or adaptive processing [Tufts], i.e., utilize information from adjacent frequency channels. Another possible approach is to exploit the psychoacoustic fact that the pitch discrimination of the auditory system is less acute at higher frequencies [Flanagan]. Therefore, rather than using the same analysis filter over the entire frequency spectrum, one might use a set of analysis filters that have bandwidths that increase with frequency and durations that decrease with frequency, thus matching the resolution of the processing system to the resolution of the ear. Such a system might be based on strategies such as constant-Q or third-octave analysis. One might also attempt to to circumvent the problem of temporal resolution vs. spectral resolution by implementing the rate changes selectively, in a feature dependent manner. Thus, the analysis filter would be conservatively designed to guarantee adequate spectral resolution but not, necessarily, to provide temporal resolution sufficient for resolving rapid

transitions. Since the rate-change system formulated in Chapter 6 is an identity system in the absence of parameter modification, one would then selectively change the rate of only the steady-state or slowly-varying portions of the speech, while leaving the rapid transitions unchanged. A set of rules could, in principle, be formulated for applying feature-dependent rate changes based on criteria such as maximizing the intelligibility of the rate-changed speech or emulating the manner in which human speakers change the rate of their own speech [Goldman-Eisler, 1961, 1968; Toong].

# REFERENCES


Agarwal, R.C., and C.S. Burrus, "Fast One-Dimensional Digital Convolution by Multidimensional Techniques," IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-22, no. 1, pp. 1-10, February 1974.

Allen, J.B. and L.R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proc. IEEE, vol. 65, no. 11, pp. 1558-1564, November 1977.

---, "Short-Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-25, no.3, pp. 235-238, June 1977.

Callahan, M.W., "Acoustic Signal Processing Based on the Short-time Spectrum," Ph.D. Thesis, Computer Sci. Dept., Univ. of Utah, Tech. Rept. no. UTEC-CSc-76-209, March 1976.

Davenport, W.B. and W.L. Root, An Introduction to the Theory of Random Signals and Noise, New York: McGraw-Hill, 1958.

Fairbanks, G., W.L. Everitt, and R.P. Jaeger, "Method for Time or Frequency Compression-Expansion of Speech," IRE Trans., Professional Group on Audio, vol. AU-2, no. 1, pp. 7-12, January-February 1954.

--- --- --- , Recording Device, Washington, D.C., May 12, 1959, U.S. Patent No. 2,886,650.

Flanagan, J.L., Speech Analysis Synthesis and Perception, 2nd ed., Berlin: Springer-Verlag, 1972.

--- and R.M. Golden, "Phase Vocoder," Bell Syst. Tech. J., vol. 45, pp. 1493-1509, November 1966.

Fletcher, H., Speech and Hearing, New York: Van Nostrand Co., 1929, pp. 293-294.

Gabor, D., "Theory of Communication," J.IEE, no. 93, Pt. III, pp. 429-441, November 1946.

Gambardella, G., "Properties of Short-Time Spectral Analysis Performed by the Peripheral Auditory System," Int'l. Congr. Cybernetics 1st., London: Gordon and Breach, 1970.

---, "A Contribution to the Theory of Short-Time Spectral Analysis with Non-uniform Bandwidth Filters," IEEE Trans. Circuit Theory, vol. CT-18, no. 4, July 1971.

Gersho, A., and N. DeClaris, "Duality Concepts in Time-Varying Linear Systems," 1964 IEEE Int'l. Conv. Record, Pt. 1, pp. 344-356.

Goldman-Eisler, "The Significance of Changes in the Rate of Articulation," Language and Speech, vol. 4, part III, June-September 1961.

---, Psycholinguistics Experiments in Spontaneous Speech, New York: Academic Press, 1968.

Heinz, J.M., and K.N. Stevens, "On the Properties of Voiceless Fricative Constants," J. Acoust. Soc. Amer., vol. 33, no. 5, pp. 589-596, May 1961. Reprinted in Readings in Acoustic Phonetics, I. Lehiste (editor), Cambridge: M.I.T. Press, 1967.

Hildebrand, F.B., Finite Difference Equations and Simulations, Englewood Cliffs: Prentice Hall, 1968.

Huggins, A.W.F., "More Temporally Segmented Speech: Is Duration or Speech Content the Critical Variable in its Loss of Intelligibility?" Quarterly Progress Report No. 114, Research Laboratory of Electronics, M.I.T., July 15, 1974, pp. 185-193.

Kharkevich, A.A., Spectra and Analysis (translated from the Russian), Consultants Bureau Enterprises, New York, 1960.

Klumpp, R.G., and J.C. Webster, "Intelligibility of Time-Compressed Speech," J. Acoust. Soc. Amer. vol. 31, pp. 265-267, 1961.

Lee, F.F., "Time Compression and Expansion of Speech by the Sampling Method," J. Audio Eng. Soc., vol. 20, no. 9, pp. 738-742, November 1972.

Moorer, J.A., "The Use of the Phase Vocoder in Computer Music Applications," 55th Convention of the Audio Engineering Soc., Preprint no. 1146 (E-1), October 1976.

Neuburg, E.P., "Simple Pitch-Dependent Algorithm for High-Quality Speech Rate-Change," 93rd Meeting Acoust. Soc. Amer., June 1977. Abstract, J. Acoust. Soc. Amer., vol. 61, suppl. no. 1, Spring 1977.

Oetken, G., T.W. Parks, and H.W. Scheussler, "New Results in the Design of Digital Interpolators," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-23, no. 3, pp. 301-309, June 1975.

Oppenheim, A.V., and R.W. Schafer, Digital Signal Processing, Englewood Cliffs: Prentice Hall, 1975.

Parsons, T.W., "Separation of Speech from Interfering Speech by Means of Harmonic Selection," J. Acoust. Soc. Amer., vol. 60, no. 4, pp. 911-918, October 1976.

Patisaul, C.R. and J.C. Hammett, Jr., "Time-Frequency Resolution Experiment in Speech Analysis and Synthesis," J. Acoust. Soc. Amer., vol. 58, no. 6, pp. 1296-1307, December 1975.

Portnoff, M.R., "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-24, no. 3, pp. 243-248, June 1976.

---, "A Mathematical Framework for Time-Scale Modification of Speech," 93rd Meeting Acoust. Soc. Amer., June 1977. Abstract, J. Acoust. Soc. Amer. vol. 61, suppl. no. 1, Spring 1977.

Schafer, R.W. and L.R. Rabiner, "Design and Simulation of A Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 165-174, June 1973a.

--- ---, "A Digital Signal Processing Approach to Interpolation," Proc. IEEE, vol. 61, no. 6, pp. 692-702, June 1973b.

--- ---, "Digital Representations of Speech Signals," Proc. IEEE, vol. 63, no. 4, pp. 662-677, April 1975.

Schroeder, M.R., "Vocoders: Analysis and Synthesis of Speech," Proc. IEEE vol. 54, pp. 720-734, May 1966.

Schweppe, F.C., Uncertain Dynamical Systems, Englewood Cliffs: Prentice Hall, 1973.

Scott, R.J., and S.E. Gerber, "Pitch Synchronous Time Compression of Speech," Proc. Conf. Speech Comm. Processing, April 1972, pp. 63-65.

Steinberg, J.D., "Acoustics, The Sense of Hearing," in Electrical Engineers Handbook, New York: John Wiley and Sons, 1936, pp. 9-32.

Stockham, T.G., Jr., "High-Speed Convolution and Correlation," AFIPS Conf. Proc., 1966, Spring Joint Computer Conf. Reprinted in Digital Signal Processing, L.R. Rabiner and C.M. Rader (editors), New York: IEEE Press, 1972.

Toong, H.D., "A Study of Time-Compressed Speech," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, M.I.T., 1974.

Tufts, D.W., "Adaptive Line Enhancement and Spectrum Analysis," Proc. IEEE vol. 65, no. 1, pp.169-170, January 1977.

Wang, R.J., "Optimum Window Length for the Measurement of Time-Varying Power Spectra," J. Acoust. Soc. Amer., vol 52, no. 1 (part 1), pp. 33-38, January 1972.

Weinstein, C.J., "Short-Time Fourier Analysis and Its Inverse," S.M. Thesis, Elect. Eng. Dept., M.I.T., 1966.

Youngberg, J.E. and S.F. Boll, "Constant-Q Signal Analysis and Synthesis," to be published Proc. 1978 IEEE Int'l. Conf. Acoust. Speech, and Signal Processing, April 1978.

Zadeh, L.A., "A General Theory of Signal Transmission Systems," J. Franklin Inst., vol. 253, pp. 293-312, April 1952.

# APPENDIX

## DERIVATIONS FOR SHORT-TIME FOURIER ANALYSIS OF UNVOICED SPEECH

### A.1 INTRODUCTION

The purpose of this appendix is to develop the results, quoted in Chapter 5, for the short-time Fourier analysis of unvoiced speech. An unvoiced speech signal, $x(n)$, is modelled as the output of a quasi-stationary linear system, $t(n,m)$, driven by a stationary Gaussian white-noise process $u(n)$. The excitation, $u(n)$ is characterized by its autocorrelation function

$$R_u(\tau) = E\{u(n+\tau)u^*(n)\} = \sigma_u^2 \delta(\tau) \qquad (A.1)$$

or, equivalently, by its power spectrum

$$S_u(\omega) = \sigma_u^2 \qquad (A.2)$$

where

$$R_u(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_u(\omega)\exp[j\omega\tau]\,d\omega \; . \qquad (A.3)$$

The unvoiced speech signal, $x(n)$, is characterized by its autocorrelation function

$$R_x(n,\tau) = E(x(n+\tau)x^*(n)) \qquad (A.4)$$

or, its time-varying power spectrum

$$S_x(n,\omega) = \sigma_u^2 |T(n,\omega)|^2 , \qquad (A.5)$$

where $R_x(n,\tau)$ and $S_x(n,\omega)$ are related by the partial Fourier transform

$$R_x(n,\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega)\exp[j\omega\tau]\,d\omega \qquad (A.6)$$

## A.2 THE MODIFIED CORRELATION FUNCTION FOR THE SHORT-TIME FOURIER TRANSFORM

The modified correlation function, $K_x(n,\omega,\tau,\epsilon)$, for the short-time Fourier transform, $X_2(n,\omega)$, is defined as the expected value

$$K_x(n,\omega,\tau,\epsilon) = E(X_2(n+\tau,\omega-\tfrac{\epsilon}{2})X_2^*(n,\omega+\tfrac{\epsilon}{2}))\exp[-j(n+\tfrac{\tau}{2})\epsilon] . \qquad (A.7)$$

$K_x(n,\omega,\tau,\epsilon)$ can be expressed in terms of $S_x(n,\omega)$, the time-varying power spectrum for $x(n)$, and the analysis filter, $H(\omega)$, by substituting the definition of the short-time Fourier transform

$$X_2(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)x(m)\exp[-j\omega m] , \qquad (A.8)$$

into the definition (A.7) to obtain

$$K_x(n,\omega,\tau,\epsilon) = E\Big(\Big(\sum_{p=-\infty}^{\infty} h(n+\tau-p)\,x(p)\exp\left[-j\left(\omega-\tfrac{\epsilon}{2}\right)p\right]\Big)$$

$$\times\ \Big(\sum_{q=-\infty}^{\infty} h(n-q)\,x(q)\exp\left[-j\left(\omega+\tfrac{\epsilon}{2}\right)q\right]\Big)^{*}\Big)\exp\left[-j\left(n+\tfrac{\tau}{2}\right)\epsilon\right]$$

$$= \sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty} E(x(p)\,x^{*}(q))\,h(n+\tau-p)\,h^{*}(n-q)$$

$$\times\ \exp\left[-j\left(\left(\omega-\tfrac{\epsilon}{2}\right)p-\left(\omega+\tfrac{\epsilon}{2}\right)q+\left(n+\tfrac{\tau}{2}\right)\epsilon\right)\right] .$$

Making the change of variables, $r = n+\tau-p$ and $s = n-q$ gives

$$K_x(n,\omega,\tau,\epsilon) = \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} E(x(n+\tau-r)\,x^{*}(n-s))\,h(r)\,h^{*}(s)$$

$$\times\ \exp\left[j\left(\left(\omega-\tfrac{\epsilon}{2}\right)r-\left(\omega+\tfrac{\epsilon}{2}\right)s-\omega\tau\right)\right]$$

and using the definition (A.4) for the autocorrelation function of $x(n)$,

$$K_x(n,\omega,\tau,\epsilon) = \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} R_x(n,\tau-r+s)\,h(r)\,h^{*}(s)\exp\left[j\left(\left(\omega-\tfrac{\epsilon}{2}\right)r-\left(\omega+\tfrac{\epsilon}{2}\right)s-\omega\tau\right)\right] .$$

Replacing $R_x(n,\tau-r+s)$ with its representation, (A.6), as the inverse partial Fourier transform of the time-varying power spectrum, $S_x(n,\omega)$, gives

$$K_x(n,\omega,\tau,\epsilon) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n,\xi) \exp[j\xi(\tau-r+s)]\, d\xi$$

$$\times\, h(r)\, h^*(s) \exp[j\{(\omega-\tfrac{\epsilon}{2})r - (\omega+\tfrac{\epsilon}{2})s - \omega\tau\}]$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n,\xi) \sum_{r=-\infty}^{\infty} h(r) \exp[-j(\xi-\omega+\tfrac{\epsilon}{2})r]$$

$$\times \sum_{s=-\infty}^{\infty} h^*(s) \exp[j(\xi-\omega-\tfrac{\epsilon}{2})s] \exp[j(\xi-\omega)\tau] \,.$$

Finally, making the change of variables $\varphi = \xi - \omega$ and carrying out the summations over r and s, gives the desired result

$$K_x(n,\omega,\tau,\epsilon) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(n,\omega+\varphi)\, H(\varphi+\tfrac{\epsilon}{2})\, H^*(\varphi-\tfrac{\epsilon}{2}) \exp[j\varphi\tau]\, d\varphi \,. \tag{A.9}$$

Note that the time-dependence (on n) of $K_x(n,\omega,\tau,\epsilon)$ is through the time-varying power spectrum $S_x(n,\omega)$. Consequently, for a stationary signal $x(n)$, $S_x(n,\omega)$ becomes $S_x(\omega)$, and the modified correlation function is independent of the absolute time n. Further, if $x(n)$ is quasi-stationary, then $S_x(n,\omega)$ and, hence, $K_x(n,\omega,\tau,\epsilon)$ are slowly-varying functions of n.

## A.3 A SECOND-ORDER APPROXIMATION FOR $K_X(n,\omega,\tau,\epsilon)$ IN $\tau$ AND $\epsilon$

An approximation to the modified correlation function, $K_X(n,\omega,\tau,\epsilon)$, can be obtained by expanding $K_X(n,\omega,\tau,\epsilon)$, given by eqn. (A.9), in a two-dimensional Taylor series in the variables $\tau$ and $\epsilon$ about the point $\tau = \epsilon = 0$. The coefficients of the expansion are obtained by assuming $\tau$ to be a continuous variable and formally differentiating the right-hand side of eqn. (A.9) with respect to $\tau$ and $\epsilon$. Thus,

$$K_X(n,\omega,\tau,\epsilon) = J_X(n,\omega)\{1 - j[M(n,\omega)\tau - m(n,\omega)\epsilon] - \tfrac{1}{2}[D^2(n,\omega)\tau^2 + 2\mu(n,\omega)\tau\epsilon + d^2(n,\omega)\epsilon^2] + \ldots\} \tag{A.10}$$

where

$$\begin{aligned}
J_X(n,\omega) &= K_X(n,\omega,0,0)\\
M(n,\omega) &= -j(\partial_\tau K_X(n,\omega,0,0)) \;/\; J_X(n,\omega)\\
m(n,\omega) &= j(\partial_\epsilon K_X(n,\omega,0,0)) \;/\; J_X(n,\omega)\\
D^2(n,\omega) &= -(\partial^2_{\tau\tau} K_X(n,\omega,0,0)) \;/\; J_X(n,\omega)\\
d^2(n,\omega) &= -(\partial^2_{\epsilon\epsilon} K_X(n,\omega,0,0)) \;/\; J_X(n,\omega)\\
\mu(n,\omega) &= -(\partial^2_{\tau\epsilon} K_X(n,\omega,0,0)) \;/\; J_X(n,\omega) = -(\partial^2_{\epsilon\tau} K_X(n,\omega,0,0)) \;/\; J_X(n,\omega)
\end{aligned} \tag{A.11}$$

and $\partial_\tau$ and $\partial_\epsilon$ denote the partial-derivative operators with respect to $\tau$ and $\epsilon$.

Because the expressions for the moments $M(n,\omega)$ $m(n,\omega)$, $D(n,\omega)$, $d(n,\omega)$, and $\mu(n,\omega)$ are still rather complicated, the expansion (A.10) offers little insight into the effects of processing unvoiced speech. These moments are simplified considerably by assuming that the bandwidth of the analysis filter, $H(\omega)$, is sufficiently narrow that the spectrum of the unvoiced speech can be considered as constant over any frequency interval less than the bandwidth of $H(\omega)$. This assumption is the usual assumption invoked in spectral analysis and is reasonable for the analysis of unvoiced speech using an analysis filter, $H(\omega)$, sufficiently narrow to resolve the harmonics of voiced speech. Thus, the first and second moments will be calculated based on the approximation

$$K_x(n,\omega,\tau,\epsilon) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(n,\omega+\varphi)\,H(\varphi+\tfrac{\epsilon}{2})\,H^*(\varphi-\tfrac{\epsilon}{2})\exp[j\varphi\tau]\,d\varphi$$

$$\approx S_x(n,\omega)\ \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\varphi+\tfrac{\epsilon}{2})\,H^*(\varphi-\tfrac{\epsilon}{2})\exp[j\varphi\tau]\,d\varphi\ . \qquad \text{(A.12)}$$

Defining the moments of the analysis filter, $h(n)$, up to second order as

$$J_h = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\omega)|^2 d\omega = \sum_{n=-\infty}^{\infty} |h(n)|^2 \qquad \text{(A.13a)}$$

$$M_h = \frac{1}{2\pi J_h}\int_{-\pi}^{\pi} \omega |H(\omega)|^2 d\omega \qquad \text{(A.13b)}$$

$$m_h = \frac{1}{J_h}\sum_{n=-\infty}^{\infty} n|h(n)|^2 \qquad \text{(A.13c)}$$

$$D_h^2 = \frac{1}{2\pi J_h} \int_{-\pi}^{\pi} \omega^2 |h(\omega)|^2 d\omega \qquad \text{(A. 13d)}$$

$$d_h^2 = \frac{1}{J_h} \sum_{n=-\infty}^{\infty} n^2 |h(n)|^2 \qquad \text{(A. 13e)}$$

and

$$\mu_h = \frac{-j}{4\pi J_h} \int_{-\pi}^{\pi} \omega [\, H'(\omega) H^*(\omega) - H(\omega) H'^*(\omega) ] d\omega \,, \qquad \text{(A. 13f)}$$

the moments of $K_x(n, \omega, \tau, \epsilon)$ with respect to $\tau$ and $\epsilon$ then become

$$J_x(n, \omega) = K_x(n, \omega, 0, 0) \qquad \text{(A. 14a)}$$

$$M(n, \omega) = -j (\partial_\tau K_x(n, \omega, 0, 0)) \;/\; J_x(n, \omega) \approx M_h \qquad \text{(A. 14b)}$$

$$m(n, \omega) = j (\partial_\epsilon K_x(n, \omega, 0, 0)) \;/\; J_x(n, \omega) \approx m_h \qquad \text{(A. 14c)}$$

$$D^2(n, \omega) = -(\partial^2_{\tau\tau} K_x(n, \omega, 0, 0)) \;/\; J_x(n, \omega) \approx D_h^2 \qquad \text{(A. 14d)}$$

$$d^2(n, \omega) = -(\partial^2_{\epsilon\epsilon} K_x(n, \omega, 0, 0)) \;/\; J_x(n, \omega) \approx d_h^2 \qquad \text{(A. 14e)}$$

and

$$\mu(n, \omega) = -(\partial^2_{\tau\epsilon} K_x(n, \omega, 0, 0)) \;/\; J_x(n, \omega) \approx \mu_h \qquad \text{(A. 14f)}$$

The first order moments, $M_h$ and $m_h$, correspond to the centers of gravity of $H(\omega)$ and $h(n)$, i.e., the center frequeny of $H(\omega)$ and "center time" of $h(n)$. For the application of short-time Fourier analysis, these moments will be assumed to be zero, i.e., $h(n)$ is assumed to be a low-pass filter with its impulse response centered about $n = 0$. Replacing the moments of $K_x(n, \omega, \tau, \epsilon)$ in the Taylor series (A.10) by their approximations, (A.14), gives

$$K_x(n,\omega,\tau,\varepsilon) = J_x(n,\omega)\{1 - \frac{1}{2}[\, D_h^2\tau^2 + 2\mu_h\tau\varepsilon + d_h^2\varepsilon^2] + \ldots\} \quad (A.15)$$

and, retaining terms up to second order yields the approximation

$$K_x(n,\omega,\tau,\varepsilon) \approx J_x(n,\omega)\{1 - \frac{1}{2}[\, D_h^2\tau^2 + 2\mu_h\tau\varepsilon + d_h^2\varepsilon^2]\} \quad (A.16)$$

## A.4 AUTOCORRELATION FUNCTION AND SPECTRUM FOR THE SYNTHESIZED SIGNAL

Let $Y_2(n,\omega)$ denote a modified short-time Fourier transform and let $y(n)$ denote the signal synthesized according to the short-time Fourier synthesis formula

$$y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} f(n-r)\, Y_2(r,\omega) \exp[j\omega n]\, d\omega \,. \quad (A.17)$$

The autocorrelation function $R_y(n,\tau)$, defined as

$$R_y(n,\tau) = E\{y(n+\tau)\, y^*(n)\} \,, \quad (A.18)$$

can be expressed in terms of the modified autocorrelation function $K_y(n,\omega,\tau,\varepsilon)$ for $Y_2(n,\omega)$, and the synthesis filter, $f(n)$, by replacing $y(n)$ in eqn. (A.18) with eqn. (A.17):

$$R_y(n,\tau) = E\{(\tfrac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{p=-\infty}^{\infty} f(n+\tau-p)\,Y_2(p,\varphi)\exp[j\varphi(n+\tau)]\,d\varphi)$$

$$\times\ (\tfrac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{q=-\infty}^{\infty} f(n-q)\,Y_2(q,\varphi)\exp[j\mathfrak{f}n]\,d\mathfrak{f})^{*}\}$$

$$= (\tfrac{1}{2\pi})^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{p}\sum_{q} f(n+\tau-p)\,f^{*}(n-q)$$

$$\times\ E\{Y_2(p,\varphi)\,Y_2^{*}(q,\mathfrak{f})\}\exp[j(\varphi-\mathfrak{f})n + \varphi\tau)]\,d\varphi\,d\mathfrak{f}\ .$$

Making the change of variables $\varphi = \omega-\frac{\epsilon}{2}$, $\mathfrak{f} = \omega+\frac{\epsilon}{2}$, and $p = q + r$ gives

$$R_y(n,\tau) = (\tfrac{1}{2\pi})^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{r}\sum_{q} f(n+\tau-q-r)\,f^{*}(n-q)$$

$$\times\ E\{Y_2(q+r,\omega-\tfrac{\epsilon}{2})\,Y_2^{*}(q,\omega+\tfrac{\epsilon}{2})\}\exp[-j(n+\tfrac{\tau}{2})\epsilon]$$

$$\times\ \exp[j\omega\tau]\,d\epsilon\,d\omega$$

$$= (\tfrac{1}{2\pi})^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{r}\sum_{q} f(n+\tau-q-r)\,f^{*}(n-q)\,K_y(q,\omega,r,\epsilon)\exp[j\omega\tau]\,d\epsilon\,d\omega. \tag{A.19}$$

or, equivalently, letting $l = n-q$ and $m = \tau-r$,

$$R_y(n,\tau) = (\tfrac{1}{2\pi})^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{l}\sum_{m} f(l+m)\,f^{*}(l)\,K_y(n-l,\omega,\tau-m,\epsilon)\exp[j\omega\tau]\,d\epsilon\,d\omega\ . \tag{A.20}$$

The time-varying power spectrum $S_y(n,\omega)$, for $y(n)$ is the partial Fourier transform of $R_y(n,\tau)$ with respect to $\tau$, i.e.,

$$S_y(n,\omega) = \sum_{\tau=-\infty}^{\infty} R_y(n,\tau)\exp[-j\omega\tau] \quad . \tag{A.21}$$

Replacing $R_y(n,\tau)$ by eqn. (A.20) gives

$$S_y(n,\omega) = \sum_{\tau} \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{q}\sum_{r} f(n+\tau-q-r)\,f^*(n-q)\,K_y(q,\varphi,r,\varepsilon)$$

$$\times \exp[j\varphi\tau]\exp[-j\omega\tau]\,d\varepsilon\,d\varphi$$

and, letting $s = n + \tau - q - r$,

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{q}\sum_{r}\sum_{s} f(s)\,f^*(n-q)\,K_y(q,\varphi,r,\varepsilon)$$

$$\times \exp[j(\omega-\varphi)(n-q-r-s)]\,d\varepsilon\,d\varphi$$

$$= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{q}\sum_{r} F(\omega-\varphi)\,f^*(n-q)\,K_y(q,\varphi,r,\varepsilon)$$

$$\times \exp[j(\omega-\varphi)(n-q-r)]\,d\varepsilon\,d\varphi$$

$$= \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{r} F(\omega-\varphi)\exp[j(\omega-\varphi)(n-r)]$$

$$\times \left\{\sum_{q} f(n-q)\,K_y^*(q,\varphi,r,\varepsilon)\exp[-j(\varphi-\omega)q]\right\}^* d\varepsilon\,d\varphi$$

(A.22)

The summation over q is recognized as the short-time Fourier transform of $K_y(q,\varphi,r,\varepsilon)$, with respect to q, using f(n) as the analysis window. Assuming that $K_y(n,\omega,\tau,\varepsilon)$ varies slowly in n, so that it is low pass and narrow band compared to $F(\omega)$, the technique of Section 2.6 for

approximating the short-time Fourier transform of a narrow-band signal can be applied to (A.22) to give

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sum_{r=-\infty}^{\infty} F(\omega-\varphi)\exp[j(\omega-\varphi)(n-r)] \times \left(F(-(\varphi-\omega))K_y^*(n,\varphi,r,\epsilon)\exp[-j(\varphi-\omega)n]\right)^* d\epsilon\, d\omega \,.$$

Thus,

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \sum_{r=-\infty}^{\infty} K_y(n,\varphi,r,\epsilon)\exp[-j(\omega-\varphi)r]\, d\epsilon\, d\varphi \,. \tag{A.23}$$

## A.5 APPROXIMATION OF $K_y$ FOR TIME-SCALE MODIFICATION

The modified short-time Fourier transform $Y_2(n,\omega)$ for synthesizing rate-changed speech is

$$\begin{aligned} Y_2(n,\omega) &= a(\beta n,\omega)\exp[j\vartheta(\beta n,\omega)/\beta] \\ &= A(\beta n,\omega)\exp[j(\alpha(\beta n,\omega) + \vartheta(\beta n,\omega)/\beta)] \end{aligned} \tag{A.24}$$

where

$$\begin{aligned} X_2(n,\omega) &= a(n,\omega)\exp[j\vartheta(n,\omega)] \\ &= A(n,\omega)\exp[j(\alpha(n,\omega) + \vartheta(n,\omega))] \\ &= A(n,\omega)\exp[j\theta(n,\omega)] \end{aligned} \tag{A.25}$$

and the functions $a(n,\omega)$, $A(n,\omega)$, $\alpha(n,\omega)$, $\vartheta(n,\omega)$, and $\theta(n,\omega)$ are defined in Chapter 4. In approximating $K_y(n,\omega,\tau,\epsilon)$, the slowly-varying phase angle

$\alpha(n,\omega)$ will be neglected so that

$$Y_2(n,\omega) = A(\beta n,\omega)\exp[j\theta(\beta n,\omega)/\beta] \tag{A.26}$$

will be taken as the modified short-time Fourier transform, and $K_y(n,\omega,\tau,\epsilon)$ becomes

$$\begin{aligned} K_y(n,\omega,\tau,\epsilon) = E\{&A(\beta(n+\tau),\omega-\tfrac{\epsilon}{2})A(\beta n,\omega+\tfrac{\epsilon}{2}) \\ &\times \exp[j(\theta(\beta(n+\tau),\omega-\tfrac{\epsilon}{2}) - \theta(\beta n,\omega+\tfrac{\epsilon}{2}))/\beta]\} \\ &\times \exp[-j(n+\tfrac{\tau}{2})\epsilon] \ . \end{aligned} \tag{A.27}$$

$K_y(n,\omega,\tau,\epsilon)$ can be approximated by expanding the right-hand side of eqn. (A.27) in a two-dimensional Taylor series about $\tau = \epsilon = 0$, and the coefficients approximated in terms of the moments (A.13) of the analysis filter, $h(n)$. To obtain the coefficients in the expansion, assume that $A(n,\omega)$ and $\theta(n,\omega)$ correspond to samples of $A(t,\omega)$ and $\theta(t,\omega)$, defined as the magnitude and phase of $X_2(t,\omega)$, where $X_2(t,\omega)$ is defined by bandlimited interpolation of $X_2(n,\omega)$. Thus,

$$\begin{aligned} K_y(t,\omega,\tau,\epsilon) = E\{&A(\beta(t+\tau),\omega-\tfrac{\epsilon}{2})A(\beta t,\omega+\tfrac{\epsilon}{2}) \\ &\times \exp[j(\theta(\beta(t+\tau),\omega-\tfrac{\epsilon}{2}) - \theta(\beta t,\omega+\tfrac{\epsilon}{2}))/\beta]\} \\ &\times \exp[-j(t+\tfrac{\tau}{2})\epsilon] \ . \end{aligned} \tag{A.28}$$

For quasi-stationary speech, assume that $K_y(t,\omega,\tau,\epsilon)$ is a slowly-varying function of t, so that

$$K_y(t,\omega,\tau,\epsilon) \approx K_y(t-\tfrac{\tau}{2},\omega,\tau,\epsilon)$$
$$= E\{A(\beta(t+\tfrac{\tau}{2}),\omega-\tfrac{\epsilon}{2})\,A(\beta(t-\tfrac{\tau}{2}),\omega+\tfrac{\epsilon}{2})$$
$$\times \exp[j(\Theta(\beta(t+\tfrac{\tau}{2}),\omega-\tfrac{\epsilon}{2}) - \Theta(\beta(t-\tfrac{\tau}{2}),\omega+\tfrac{\epsilon}{2}))/\beta])$$
$$\times \exp[-j_o t] \quad . \qquad \text{(A.29)}$$

Although this approximation is not necessary, it does simplify the mathematics. For convenience, the time-scaled function

$$K_y(t/\beta,\omega,\tau/\beta,\epsilon) = E\{A(t+\tfrac{\tau}{2},\omega-\tfrac{\epsilon}{2})\,A(t-\tfrac{\tau}{2},\omega+\tfrac{\epsilon}{2})$$
$$\times \exp[j(\Theta(t+\tfrac{\tau}{2},\omega-\tfrac{\epsilon}{2}) - \Theta(t-\tfrac{\tau}{2},\omega+\tfrac{\epsilon}{2}))/\beta])$$
$$\times \exp[-j_o t/\beta] \quad . \qquad \text{(A.30)}$$

will be expanded, rather than (A.29).

Expanding the right-hand side of eqn. (A.30) in a two-dimensional Taylor series in $\tau$ and $\epsilon$ and retaining terms up to second order gives

$$K_y(t/\beta,\omega,\tau/\beta,\epsilon) = E\{A^2\} + j[\,E\{\Theta_t A^2\}\tau/\beta - E\{(\Theta_\omega+t)A^2\}\epsilon/\beta]$$
$$+ \tfrac{1}{2}[\,\tfrac{1}{2}\cdot E\{\hat{A}_{tt}A^2\} - E\{\Theta_t^2 A^2\}]\tau^2/\beta^2$$
$$- [\,\tfrac{1}{2}\cdot E\{\hat{A}_{t\omega}A^2\} - E\{\Theta_t(\Theta_\omega+t)A^2\}]\tau\epsilon/\beta^2$$
$$+ \tfrac{1}{2}[\,\tfrac{1}{2}\cdot E\{\hat{A}_{\omega\omega}A^2\} - E\{(\Theta_\omega+t)^2A^2\}]\epsilon^2/\beta^2$$
$$+ \ldots \qquad \text{(A.31)}$$

where $\hat{A} = \log A(t,\omega)$ and the subscripts t and $\omega$ denote partial differentiation of the subscripted quantity with respect to the subscript, e.g., $\Theta_t = \partial_t\Theta(t,\omega)$. Similarly, expanding

$$K_x(t,\omega,\tau,\epsilon) = E\{A(t+\tfrac{\tau}{2},\omega-\tfrac{\epsilon}{2})\,A(t-\tfrac{\tau}{2},\omega+\tfrac{\epsilon}{2}) \times \exp[j(\theta(t+\tfrac{\tau}{2},\omega-\tfrac{\epsilon}{2}) - \theta(t-\tfrac{\tau}{2},\omega+\tfrac{\epsilon}{2}))]\} \times \exp[-j\epsilon t]$$

gives

$$\begin{aligned} K_x(t,\omega,\tau,\epsilon) = {} & E\{A^2\} + j[\,E\{\theta_t A^2\}\tau - E\{(\theta_\omega+t)A^2\}\epsilon\,] \\ & + \tfrac{1}{2}[\,\tfrac{1}{2}\cdot E\{\hat{A}_{tt}A^2\} - E\{\theta_t^2 A^2\}]\tau^2 \\ & - [\,\tfrac{1}{2}\cdot E\{\hat{A}_{t\omega}A^2\} - E\{\theta_t(\theta_\omega+t)^2A^2\}]\tau\epsilon \\ & + \tfrac{1}{2}[\,\tfrac{1}{2}\cdot E\{\hat{A}_{\omega\omega}A^2\} - E\{(\theta_\omega+t)^2A^2\}]\epsilon^2 \\ & + \ldots \end{aligned} \tag{A.32}$$

which corresponds to the expansion (A.31) with $\beta = 1$. Thus,

$$K_y(t/\beta,\omega,\tau/\beta,\epsilon) = K_x(t,\omega,\tau,\epsilon) + [\,K_y(t/\beta,\omega,\tau/\beta,\epsilon) - K_x(t,\omega,\tau,\epsilon)]$$

$$\begin{aligned} = {} & K_x(t,\omega,\tau,\epsilon) \\ & + j(1/\beta - 1)[\,E\{\theta_t A^2\}\tau - E\{(\theta_\omega+t)A^2\}\epsilon\,] \\ & - \tfrac{1}{2}(1/\beta^2 - 1)[\,E\{\theta_t^2A^2\}\tau^2 - 2E\{\theta_t(\theta_\omega+t)A^2\}\tau\epsilon \\ & \qquad + E\{(\theta_\omega+t)^2A^2\}\epsilon^2\,] + \ldots \end{aligned} \tag{A.33}$$

Since the expansion for $K_x(t,\omega,\tau,\epsilon)$, to second order has already been determined, the expansion for $K_y(\frac{t}{\beta},\omega,\frac{\tau}{\beta},\epsilon)$ requires evaluating the expectations

$$\begin{aligned} & E\{\theta_t A^2\}, \qquad E\{(\theta_\omega+t)A^2\} \\ & E\{\theta_t^2A^2\}, \qquad E\{\theta_t(\theta_\omega+t)A^2\}, \quad E\{(\theta_\omega+t)^2A^2\}\,. \end{aligned} \tag{A.34}$$

Assuming Gaussian statistics for the unvoiced speech signal, $x(n)$, a tedious calculation yields

$$\mathbf{E}(\Theta_t A^2) \approx J_x(t,\omega) M_h \qquad \text{(A.35a)}$$

$$\mathbf{E}((\Theta_\omega + t) A^2)\} \approx J_x(t,\omega) m_h \qquad \text{(A.35b)}$$

$$\mathbf{E}((\Theta_t - M_h)^2 A^2\} \approx \tfrac{1}{2} J_x(t,\omega) D_h^2 \qquad \text{(A.35c)}$$

$$\mathbf{E}((\Theta_t - M_h)(\Theta_\omega + t - m_h) A^2\} \approx -\tfrac{1}{2} J_x(t,\omega) \mu_h^2 \qquad \text{(A.35d)}$$

$$\mathbf{E}((\Theta_\omega + t - m_h)^2\} A^2 \approx \tfrac{1}{2} J_x(t,\omega) d_h^2 \qquad \text{(A.35e)}$$

where the functions on the right-hand side are defined by eqns. (A.13).

Briefly, the procedure for calculating the expectations (A.34) is the following. Define $U(t,\omega)$ and $V(t,\omega)$ as the real and imaginary parts of $X_2(t,\omega)$, so that

$$X_2(t,\omega) = U(t,\omega) + jV(t,\omega)$$

and

$$A^2(t,\omega) = U^2(t,\omega) + V^2(t,\omega) ,$$

and let the subscripts $\mu$ and $\nu$ denote partial differentiation with respect to either $t$ or $\omega$. The quantity $\Theta_\mu(t,\omega)$ can be expressed as

$$\Theta_\mu(t,\omega) = \mathrm{Im}\{\partial_\mu \log X_2(t,\omega)\}$$

$$= \{(UV_\mu - U_\mu V)/A^2\} \qquad \text{for } \mu = t, \omega .$$

Further, it follows that

$$E(\theta_\mu A^2) = E(UV_\mu - U_\mu V) \tag{A.36}$$

and

$$E(\theta_\mu \theta_\nu A^2) = E((UV_\mu - U_\mu V)\cdot(UV_\nu - U_\nu V)/A^2) \tag{A.37}$$

for $\mu, \nu = t, \omega$

The expectations (A.36) and (A.37) can be evaluated from the conditional means

$$E(U_\mu | U, V) \text{ and } E(V_\mu | U, V) \tag{A.38}$$

and the conditional correlations

$$E(U_\mu U_\nu | U, V),\ E(U_\mu V_\nu | U, V),\ \text{and } E(V_\mu V_\nu | U, V) \tag{A.39}$$

Assuming Gaussian statistics for the unvoiced speech signal, the functions U, V, and their partial derivatives are jointly Gaussian. Therefore, the conditional means (A.38) and the conditional correlations (A.39) can be expressed in terms of the (unconditional) means and correlations of U, V, and their partial derivatives (see for example, [Schweppe]).

The correlation functions $E(U(t_1,\omega_1)U(t_2,\omega_2))$, $E(U(t_1,\omega_1)V(t_2,\omega_2))$, and $E(V(t_1,\omega_1)V(t_2,\omega_2))$ can be calculated by interpreting the short-time Fourier transform (for each value of $\omega$) as a demodulated narrow-band random process and paralleling the discussion in [Davenport and Root, 1958] for narrow-band Gaussian random processes. The correlation functions for the derivatives are then obtained by appropriately differentiating the correlation functions for U and V.

The desired approximation to $K_y(n,\omega,\tau,\epsilon)$ is now obtained by replacing the expectations in the series (A.33) with the values (A.35) and setting $M_h = m_h = 0$ to give

$$K_y(t/\beta,\omega,\tau/\beta,\epsilon) = K_x(t,\omega,\tau,\epsilon) - \tfrac{1}{4}(1/\beta^2 - 1)J_x(t,\omega)\,[\,D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] + \ldots . \tag{A.40}$$

Expanding $K_x(t,\omega,\tau,\epsilon)$, according to (A.15), collecting like terms, and setting $M_h = m_h = 0$ yields

$$K_y(t/\beta,\omega,\tau/\beta,\epsilon) = J_x(t,\omega)\{1 - \tfrac{1}{4}(1/\beta^2 + 1)\,[\,D_h^2\tau^2 + 2\mu_h\tau\epsilon + d_h^2\epsilon^2] + \ldots\} . \tag{A.41}$$

Scaling t and $\tau$ by $\beta$, eqn. (A.41) becomes

$$K_y(t,\omega,\tau,\epsilon) = J_x(\beta t,\omega)\{1 - \tfrac{1}{4}(1+\beta^2)\,[\,D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2] + \ldots\} \tag{A.42}$$

and defining the parameter $\gamma$ such that

$$\gamma^2 = \tfrac{1}{2}(1+\beta^2) \tag{A.43}$$

gives

$$K_y(t,\omega,\tau,\epsilon) = J_x(\beta t,\omega)\{1 - \tfrac{\gamma^2}{2}\,[\,D_h^2\tau^2 + 2\mu_h\tau(\epsilon/\beta) + d_h^2(\epsilon/\beta)^2] + \ldots\} \tag{A.44}$$

Finally, replacing t with n and comparing the resulting expansion for $Ky(n,\omega,\tau,\epsilon)$ with the expansion (A.15), for $K_x(n,\omega,\tau,\epsilon)$ shows that

$$K_y(n,\omega,\tau,\epsilon) \approx K_x(\beta n,\omega,\gamma\tau,\gamma\epsilon/\beta) \qquad \text{(A.45)}$$

## A.6 POWER SPECTRUM FOR SYNTHETIC RATE-CHANGED UNVOICED SPEECH

The time-varying power spectrum for the synthesized rate-changed unvoiced speech signal, $y(n)$, is obtained by substituting the approximation, (A.45), for the modified correlation function, $K_y(n,\omega,\tau,\epsilon)$, into the expression, (A.23), for the time-varying power spectrum for the synthesized signal. Thus,

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \sum_r K_x(\beta n,\varphi,\gamma r,\gamma\epsilon/\beta) \times \exp[-jr(\omega-\varphi)]\,d\epsilon\,d\varphi \; . \qquad \text{(A.46)}$$

Substituting the expression (A.9) for $K_x(n,\omega,\tau,\epsilon)$ gives

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \times \sum_r \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\beta n,\varphi+\xi) H(\xi+\gamma\epsilon/2\beta) H^*(\xi-\gamma\epsilon/2\beta) \times \exp[j\gamma\xi r]\,d\xi \exp[-jr(\omega-\varphi)]\,d\epsilon\,d\varphi \; . \qquad \text{(A.47)}$$

Defining the function $H_1(\omega)$ as

$$H_1(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\omega+\tfrac{\epsilon}{2}) H^*(\omega-\tfrac{\epsilon}{2})\,d\epsilon$$

$$= \frac{1}{\pi}\int_{-\pi}^{\pi} H(2\omega-\psi) H^*(\psi)\,d\psi \qquad \text{(A.48)}$$

gives

$$S_y(n,\omega) = \left(\frac{1}{2\pi}\right)^2 \int_{-\pi}^{\pi} |F(\omega-\varphi)|^2 \int_{-\pi}^{\pi} S_x(\beta n, \varphi+\zeta) H_1(\zeta) \frac{\beta}{\gamma} d\zeta$$

$$\times \sum_r \exp[-jr(\omega-\varphi-\gamma\zeta)]\, d\varphi$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} |F(\gamma\zeta)|^2 S_x(\beta n, \omega-\gamma\zeta+\zeta) H_1(\zeta) \frac{\beta}{\gamma} d\zeta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} |F(\gamma\zeta)|^2 H_1(\zeta) S_x(\beta n, \omega+(1-\gamma)\zeta) \frac{\beta}{\gamma} d\zeta \ .$$

Making the change of variables $\varphi = (1-\gamma)\zeta$ gives

$$S_y(n,\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |F(\gamma\varphi/(1-\gamma))|^2 H_1(\varphi/(1-\gamma)) S_x(\beta n, \omega+\varphi)\, d\varphi \cdot \beta/\gamma(1-\gamma) \tag{A.49}$$

and defining

$$G_\beta(\omega) = [\beta/\gamma(1-\gamma)] \cdot |F(\omega\gamma/(1-\gamma))|^2 H_1(\omega/(1-\gamma)) \tag{A.50}$$

gives the desired result for the time-varying power spectrum of the synthetic rate-changed speech, i.e.,

$$S_y(n,\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\beta n, \omega+\varphi) G_\beta(\varphi)\, d\varphi \ . \tag{A.51}$$

## BIOGRAPHICAL NOTE

Michael Rodney Portnoff was born on July 1, 1949 in Newark, New Jersey. Until comming to M.I.T., he resided and attended public school in West Orange, N.J. As an undergraduate at M.I.T. he was enrolled in the Course VI-2 (honors) program in Electrical Science and Engineering and the Course VI-A cooperative program with Bell Telephone Laboratories, Murray Hill, N.J. In June 1973, he simultaneously received the S.B., S.M., and E.E. degrees.

As a graduate student at M.I.T., he has been a teaching assistant for undergraduate and graduate courses in signals and systems and for special courses in digital signal processing at the M.I.T. Center for Advanced Engineering Study. He has also been a research assistant at the M.I.T. Research Laboratory of Electronics. His research interests are in the theory of digital signal processing and its application to speech, picture, seismic, and biomedical signal processing.

Mr. Portnoff is a member of Eta Kappa Nu, Tau Beta Pi, Sigma Xi, and the Institute of Electrical and Electronics Engineers. He was the recipient of the 1977 I.E.E.E. Browder J. Thompson Memorial Prize Award for the best paper appearing in any I.E.E.E. publication by an author under 30 years of age.